STRUCTURE AND EXPRESSION OF THE GENE ENCODING OVINE β-LACTOGLOBULIN

S. ALI

Thesis submitted for the degree of

Doctor of Philosophy

University of Edinburgh

1989

Abstract

β-lactoglobulin is the major whey protein in the milk of ruminants. It is an 18,300 dalton polypeptide whose function in milk is unclear. The detailed characterisation of two very similar but non-identical genomic clones encoding ovine β-lactoglobulin is presented in this thesis, including mapping of several repeats. DNA sequencing of the entire transcription unit and 1.9 kb of 3' flanking region was carried out. These sequences and 810 bp of 5' flanking sequences are sufficient for correct regulation of the ovine β-lactoglobulin gene in transgenic mice (S. Harris (this lab), unpublished results).

The β-lactoglobulin gene was shown to be mammary-specific and its expression during pregnancy was followed in sheep. The pattern of expression of the β-lactoglobulin gene has been compared with patterns of expression of other milk protein genes. β-lactoglobulin, α_{s1}- and β-casein genes are coordinately expressed during pregnancy. Their mRNA levels rose from day 100 of gestation, following the first observation of exudate in histological sections (at day 90). α-lactalbumin, α_{s2}- and κ-casein mRNAs were first evident at or near parturition. Thus, the β-lactoglobulin gene is temporally and tissue-specifically regulated. Differences in β-lactoglobulin gene CpG methylation between mammary gland and liver were also demonstrated.

The gene is 4.9 kb long and contains seven exons. It codes for a 180 amino-acid polypeptide, containing an 18 residue signal peptide. Translation starts in exon I and terminates in exon VI. Exon VII is entirely non-coding. S1 analysis mapped the major transcriptional start site 34 bp downstream of the TATA-box. The gene is extremely G+C-rich (~60%) throughout its length, the G+C-rich domain extending over at least 13 kb. Computer analyses of these sequences for the presence of transcriptionally important regions, are presented. Furthermore, sequence comparisons show that one of the repeat regions in the ovine β-lactoglobulin gene, present just 3' of the final exon, is similar to a ruminant repeat family previously described.

DNA sequence analysis has also shown that the two different genomic clones encode ovine β-lactoglobulin variants A and B, respectively. The existence of at least five haplotypes is demonstrated.

The genes encoding rodent urinary protein, α1-acid glycoprotein, serum retinol binding protein and apolipoprotein-D have a similar organisation of exons

and introns to the β-lactoglobulin gene. In particular, a comparison between β-lactoglobulin and serum retinol binding protein shows that both genes encode equivalent elements of three-dimensional protein structure within analogous exons. These proteins are all members of a large, diverse family of secretory proteins, many of which function in binding small lipophilic molecules. The evolutionary relationship between the different proteins is analysed in detail using tree-building, gene and protein structure comparisons. Possible relationship of this family of proteins with a family of intracellular, lipophilic molecule binding proteins is discussed.

Publications

Some of this work has already been published,

"Characterisation of the gene encoding ovine beta-lactoglobulin: similarity to the genes for retinol binding protein and other secretory proteins" by Ali, S. and Clark, A. J. (1988). J. Mol. Biol. 199, 415-426.

"Complete nucleotide sequence of the genomic ovine beta-lactoglobulin gene" by Harris, S., Ali, S., Anderson, S., Archibald, A. L. and Clark, A. J. (1988). Nucleic Acids Research 16, 10379-10380.

Reprints of these publications are bound at the back of the thesis.

Acknowledgments

I would like to thank first and foremost my supervisor Dr. John Clark, without whose guidance this work would not have progressed. My thanks also to Dr. Rick Lathe for offering me the studentship and his subsequent interest. Thanks also to Dr. John Bishop for his supervision (I regret not having sought his advice more often than I did). I am also extremely grateful to Drs. Alan Archibald, Stephen Harris, Maggie McClenaghan, Paul Simons and Bruce Whitelaw for their interest and their help with matters technical and otherwise. I am grateful for their willingness to let me discuss their results in this thesis. My thanks also to Jen Anderson, John Bowman, Morag Robertson, Wendy Shepherd and John Webster. Sympathies are extended to them and particularly to cell mates Drs. Pamela Brown and Jean-Luc Vilotte for having had to put up with me.

I would also like to thank the people who shared their results with me, particularly for divulging their unpublished results. I am especially grateful to Dr Lindsay Sawyer for the many discussions.

My thanks also to everybody at IAPGR in Edinburgh and in particular David Drury, Bill Seawright, George Davidson and Alaistair McGregor for their work with the sheep. Thanks also to Norman Russell (Roslin Photography) and to Sir Frank (in the Genetic Dept) for his excellent and speedy work.

This work was done at AFRC-IAPGR on an AFRC studentship.

DEDICATION

to family and friend

Abbreviations

bp	base pairs
dal(s)	daltons
DNA	deoxyribonucleic acid
DNasel	deoxyribonuclease I
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
kb	kilobases (1000 x bp)
kdal	kilodaltons (1000 x daltons)
A_n	optical density $_{\text{wavelength in nanometers}}$
PEG	polyethylene glycol
RNaseA	ribonuclease A
SDS	sodium dodecyl sulphate
Tris	tris (hydroxymethyl) aminomethane
TEMED	N, N, N', N' tetramethylethylenediamine
dNTP	2'-deoxy (N) 5'-triphosphate (N = adenosine, cytidine, guanosine)
dTTP	thymidine 5'-triphosphate
ddNTP	2', 3'-dideoxy (N) 5'-triphosphate (N = adenosine, cytidine, guanosine or thymidine)
MW	molecular weight

poly (A)	polyriboadenylic acid
rATP	adenosine 5'-triphosphate
RNA	ribonucleic acid
w/v	weight/volume

g	gram
mg	milligram
µg	microgram
ng	nanogram
pg	picogram
l	litre
ml	millilitre
µl	microlitre
M	molar
mM	millimolar
µM	micromolar

Table of Contents

List of figures

List of Tables

Chapter I: INTRODUCTION

The mammary gland is unique to mammals, providing nutrition to the young. It is unusual in being one of the few tissues which is poorly developed at birth and which undergo major changes after birth. It is under the control of many hormones, as well as local interactions. In the adult, the mammary gland undergoes cycles of growth, secretion of milk (lactation) and regression. Since a major part of mammary gland development occurs in the adult, hormonal regulation can be more readily analysed than in other tissues where most development is complete by birth and embryos and fetuses need to be used. Furthermore, although milk consists of thousands of components, some are present at high levels. In particular, most of the milk proteins consist of a few major species synthesised in the mammary gland. The genes encoding these proteins are present as single copy genes which are expressed at high levels and can be used to follow mammary gland development. Thus, the mammary gland offers a unique system for the study of hormonal regulation and induction of tissue-specific genes. Moreover, these genes are expressed at high levels during pregnancy and lactation, followed by a dramatic reduction in expression (or switch off of expression) during mammary gland involution and subsequent reinduction, possibly at a much later stage.

The mammary gland is also important for study because of its importance in the mammalian life cycle. Milk provides the young with all the nutrition it requires. Many of the components of milk have evolved to fulfil important roles in nutrition and development of the young. Thus, the milk protein α-lactalbumin is required for the synthesis of the principal sugar in milk, lactose. These and other milk constituents have been extensively studied. Molecular biology can also provide important

information about the structure and function(s) of these components.

This thesis analyses the gene structure of the ovine milk protein gene encoding β-lactoglobulin (BLG), discusses its evolution and possible function(s) of the protein. Some of the work presented describes patterns of expression of the BLG gene during pregnancy and lactation in sheep and compares its expression with that of other milk protein genes. The mammary gland, milk and its components are introduced in this chapter. Since a vast body of work has been carried out on milk and the mammary gland, only information relevant to the work presented is described, together with additional information where it helps to give a more complete picture.

1.1 MILK COMPOSITION AND THE STRUCTURE AND FUNCTION OF MILK PROTEINS

Milk is a heterogeneous substance containing "on the order of 100,000 different molecular species" (Walstra and Jenness, 1984), providing all of the nutrients required for the early development of the young. The composition of milk can vary between species, although all milks contain similar components. Seal milk, for example, contains much more fat, and much less lactose, than cows' milk (see Hambraeus (1982) for review). The information presented below has been obtained from work done on bovine milk, unless otherwise stated. Sheep milk composition is similar to that of bovine milk. For references on these topics see the reviews by Swaisgood, Schmidt, Jenness and Mepham et al., in Developments in Dairy Chemistry-1 (1982); and the book by Walstra and Jenness (1984)).

Table 1.1 shows the composition of bovine milk. The major components are

water, fats, protein and lactose. Lactose is the only sugar present in milk, other carbohydrates being present only in trace amounts. Lactose concentrations appear to be important for the osmotic balance of milk and initiation of its synthesis has been correlated with the start of milk secretion. The majority of milk fat is found in the form of fat globules and much of the milk protein is casein, which forms "micelles" (see section 1.1.1).

The fat globules are large spherical particles, 0.1-10 μm in diameter, surrounded by a membrane derived from the mammary secretory cells, with a composition similar to that of the outer cell membrane of mammary epithelial cells. They contain most of the milk fat, mono-, di- and tri-glycerides and fatty acids, as well as sterols, corticosteroids and vitamins A, D, E and K. The fat globule membrane contains some membrane proteins, lipids and sterols. The membrane function appears to be to prevent fat globules from flocculating and coalescing, and also protect them against enzyme action. Much smaller, 10 nm diameter, lipoprotein particles are also present in milk. They make up about 0.4 % of the total milk fat and are not surrounded by a membrane.

Cells are also found to be present in milk. These are generally leukocytes, about 100,000 per ml of milk. Their number goes up in diseased mammary glands, for example in mastitic cows. Cell debris, bits of membrane, etc., are also present in milk, particularly towards the end of lactation.

The non-particulate serum contains water, lactose (in cow 46 g/kg of lactose is present and about 0.1 g of other carbohydrates), minerals (calcium, magnesium, potassium, sodium, chloride, phosphate, sulphate and bicarbonate ions being the most abundant), trace elements (for example zinc iron, copper), organic acids (such as citrate), some lipid and vitamins, enzymes (such as lactoperoxidase and acid phosphatase) and many other trace substances. The milk serum also contains

mammary-specific (β-lactoglobulin, α-lactalbumin, lactoferrin and a little casein) and blood serum-derived proteins (such as serum albumin), as well as free amino-acids and peptides (table 1.1) (see Walstra and Jenness, 1984).

1.1.1 Milk Proteins

Bovine milk contains about 30-35 g of protein per litre of milk. These proteins can be divided into two classes, the caseins and the "serum proteins", according to solubilities. Lowering the pH of milk below pH 4.7 causes precipitation of the caseins, whilst the so-called serum proteins remain in solution. The caseins make up about 80% of bovine milk protein (about 26 g/litre), aggregated to form micelles. The serum proteins are free in solution and do not aggregate in any way. Some protein is also associated with the fat globules, as mentioned above.

In cows (and sheep) four major proteins constitute the caseins, α_{s1}-, α_{s2}-, β- and κ-caseins (see table 1.2). The serum, or "whey" proteins, consist mainly of β-lactoglobulin (54% of total whey protein), α-lactalbumin (21%) and lactoferrin, as well as serum albumin, immunoglobulins and proteose peptones.

α_{s1}-, α_{s2}- and β-caseins are about 24 kdal in size, κ-casein is 19 kdal in size. They contain relatively large numbers of proline residues, which do not participate in the formation of α-helical or β-sheet structures. The caseins therefore, contain little ordered structure. Cysteine residues are only present in α_{s2}- and κ-caseins, so disulphide bond formation does not contribute greatly to casein structure.

The caseins are highly phosphorylated. In all cases (but one) the

Table 1.1 Composition and Structure of Milk
Average quantities in 1 kg of milk

Fat Globule		
Glycerides		
triglycerides	38	g
diglycerides	0.1	g
monoglycerides	10	mg
Fatty Acids	25	mg
Sterols	100	mg
Carotenoids	0.4	mg
Vitamins A, D, E, K	2	mg
Water	60	mg
Others	30	mg
Fat Globule Membrane		
Water	80	mg ?
Protein	350	mg
Lipids		
phospholipids	210	mg
cerebrosides	30	mg
gangliosides	5	mg
neutral glycerides	+	
sterols	15	mg
Enzymes		
alkaline phosphatase	+	
xanthine oxidase	+	
others		
Copper	4	µg
Iron	100	µg
Casein Micelle		
Protein		
casein	26	g
proteose peptone	0.4	g
Salts		
calcium	800	mg
phosphate	950	mg
citrate	140	mg
magnesium, potassium, sodium, zinc, etc.	150	mg
Enzymes		
lipoprotein lipase	+	
plasmin	+	
Water		
Lipoprotein particle		
Polar lipids		
Protein		
Enzymes		
Leukocytes		

Serum		
Water	870	g
Carbohydrates		
lactose	46	g
others	0.1	g ?
Minerals		
calcium	370	mg
magnesium	75	mg
potassium	1340	mg
sodium	460	mg
chloride	1060	mg
phosphate	1080	mg
sulphate	100	mg
bicarbonate	100	mg
Trace elements		
zinc	400	µg
iron	100	µg
copper	20	µg
others		
Organic acids		
citrate	1600	mg
formate	40	mg
acetate	30	mg
lactate	30	mg
oxalate	20	mg
others	20	mg
Gases		
oxygen	6	mg
nitrogen	15	mg
Lipids		
neutral glycerides	+	
fatty acids	15	mg
phospholipids	110	mg
cerebrosides	10	mg
sterols	15	mg
others		
Vitamins		
B vitamins	200	mg
ascorbic acid	20	mg
Proteins		
caseins	+	
β-lactoglobulin	3200	mg
α-lactalbumin	1200	mg
serum albumin	400	mg
immunoglobulins	750	mg
proteose peptone	200	mg
others	400	mg
Nonprotein nitrogenous compounds		
urea	300	mg
peptides	200	mg
amino acids	300	mg
others		
Phosphoric esters	300	mg
Enzymes		
lactoperoxidase	+	
acid phosphatase	+	
many others		
Alcohol	3	mg

Adapted from Walstra and Jenness, 1984.

Table 1.2 Protein Composition of Milk from a number of Species.

	Concentration in milk (g/litre)			
	Cow	**Sheep**	**Mouse**	**Human**
Caseins			NDA	
αs1-casein	10	12		}0.4
αs2-casein	3.4	3.8		
β-casein	10	16		3
κ-casein	3.9	4.6		1
Major whey proteins				
α-lactalbumin	1	0.8	trace	1.6
β-lactoglobulin	3	2.8	none	none
Whey acidic protein	none	none	2	none
Other whey proteins				
Lactoferrin	0.1	NDA	NDA	1.4
Serum albumin	0.4	NDA	NDA	0.4
Lysozyme	trace	NDA	NDA	0.4
Immunoglobulins	0.1	NDA	NDA	1.4
Proteose peptones	0.2	NDA	NDA	NDA

NDA - No Data Available.
Taken from Clark et al.(1986) and Walstra and Jenness (1984).

Table 1.3 Immunoglobulin Composition of Various Milks.

	Concentration in milk (g/litre)					
	Cow		**Pig**		**Human**	
	Colostrum	**Milk**	**Colostrum**	**Milk**	**Colostrum**	**Milk**
IgA	3.9	0.14	10.7	7.7	17.4	1.0
IgG	-	-	58.7	3.0	0.4	0.04
IgG1	47.6	0.6	-	-	-	-
IgG2	2.9	0.02	-	-	-	-
IgM	4.2	0.05	3.2	0.3	1.6	0.1
FSC	0.2	0.05	-	-	2.1	+

- Not Known.
+ Present in very low amounts.
Redrawn from Jenness, 1982.

phosphorylated residue is a serine. Often, series of phosphorylated serines are present (for example, α_{s2}-casein contains two runs of three phosphoserines (SerP-SerP-SerP)). κ-casein contains only one phosphoserine whereas α_{s1}-, α_{s2}- and β-caseins contain 8, 10-11 and 5 phosphoserines, respectively. In addition to its low phosphorylation status κ-casein is the only casein which is glycosylated, although the extent of κ-casein glycosylation is variable.

As stated above, casein molecules contain very little ordered structure. Furthermore, α_{s1}- and α_{s2}-caseins show even distribution of hydrophobic and hydrophilic residues, with a high degree of phosphorylation. β-casein contains fewer phosphates and shows clustering of hydrophobic and hydrophilic residues, forming detergent-like, head-tail molecules.

α_{s1}-, α_{s2}- and β-caseins bind significant numbers of divalent metal ions (mostly calcium ions) and aggregate. α_{s1}-casein can bind up to 20 calcium ions per molecule of protein. Under normal concentrations of calcium ions caseins bind up to 10 moles of calcium ions per mole of protein (see Dalgleish, 1982). At least some of these calcium ions bind to the phosphoserines. Binding to aspartic and glutamic acid residues can occur but only at high concentrations of calcium. Under normal conditions in milk, the extent of calcium ion binding is not likely to exceed the number of phosphoserine residues. Binding of calcium ions reduces the charge balance which keeps caseins soluble and would lead to their precipitation (they are known as the calcium-sensitive caseins due to their insolubility in the presence of low concentrations of calcium). The reduction in charge apparently reduces electrostatic repulsion between casein molecules, allowing aggregation. Aggregation would lead to precipitation of caseins in the absence κ-casein. κ-casein is soluble in high concentrations of calcium. It associates with the other caseins and stabilises aggregates of α_{s1}-, α_{s2}- and

β-caseins.

In fresh milk, the caseins are largely present in micelles, complexes of individual molecules of the different caseins, calcium ions, calcium phosphate and calcium citrate (figure 1.1). Micelles are usually 30-300 nm in diameter, with molecular masses of 10^7-10^9 daltons, assembled from subunits of 10-15 nm diameter and about $6x10^5$ daltons. The size of micelles appears to be correlated with κ-casein concentration, suggesting its involvement on the surface of the micelle. κ-casein has a polarised structure, with a hydrophobic N-terminal region and a hydrophilic C-terminal region. It is thought that the hydrophobic region interacts with the caseins, whilst the hydrophilic region protrudes out into the aqueous surroundings. The presence of κ-casein on the surface of micelles has been confirmed by limited protease digestion of micelles (although it does not appear to be limited to the surface). Its presence on the surface of the micelle also acts to prevent micelles from flocculating together.

Current models of micelle structure suggest that the subunits are made up of all four caseins. The subunits appear to consist mainly of α_{s1}-, α_{s2}- and β-caseins, with κ-casein being present in greater amounts in subunits at micelle surface. Micelle subunits interact largely through calcium phosphate and calcium citrate. Interactions between phosphoserines of different micelle subunit caseins occur through calcium phosphate complexes (figure 1.1a).

α_{s1}-, α_{s2}- and β-caseins are hydrolysed by milk proteases (plasmin), after secretion. β-casein is digested to give γ-caseins and "proteose peptones". The γ-caseins remain associated with the casein fraction, but the proteose peptones, which have a hydrophilic structure, are present in the whey fraction. α_{s1}-casein is hydrolysed to λ-caseins and other peptides. α_{s2}-casein is also hydrolysed to give several peptides.

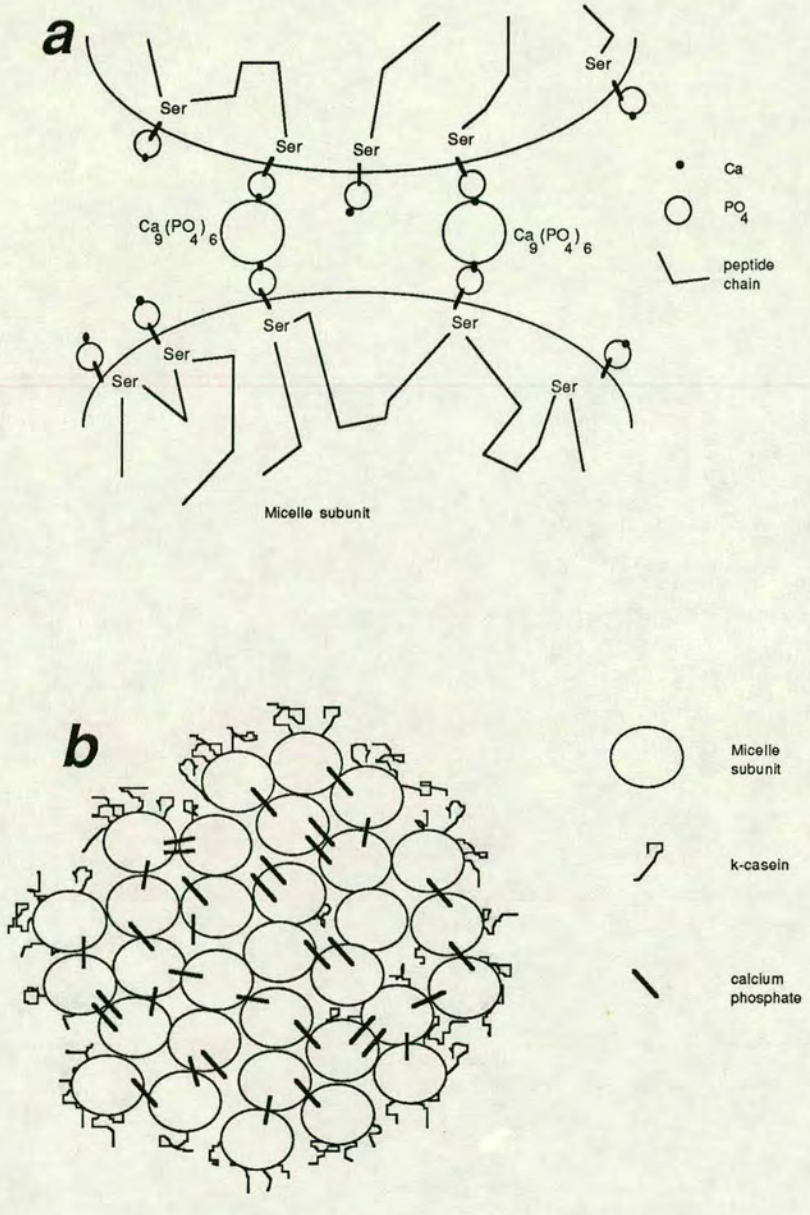


Figure 1.1 Structure of the casein micelle. The micelle subunits interact through calcium-phosphate interactions, as shown in a. The gross micelle structure is shown in b. k-casein appears to be particularly abundant on the micelle surface, although it is also present in internal subunits. The polar part of the k-casein "sticks" out into the aqueous medium, and also prevents micellar aggregation, by repulsion (redrawn from Schmidt (1982), Walstra and Jenness, 1984).

κ-casein is not hydrolysed by plasmin but by chymosin (also known as renin). Chymosin digestion of κ-casein gives "para-κ-casein", which contains the hydrophobic N-terminal region and an acidic, soluble macropeptide, the "caseinomacropeptide". Chymosin proteolysis is the first step in the breakdown of casein micelles in the stomach of the young. Cleavage of κ-casein removes the protruding hydrophilic tail, reducing repulsion between micelles and allowing flocculation. It has been suggested that κ-casein is evolutionally related to γ-fibrinogen, a protein involved in blood clotting (Jolles et al., 1978; and references therein). The possibility of an analogous function for κ-casein in milk is raised. Doolittle (1981) states, however, that the amino-acid sequence similarity is not statistically significant.

Caseins provide the majority of nitrogen for the young (making up 84% of sheep milk protein). The formation of the casein micelles allows high concentrations of the fairly insoluble calcium phosphate to be kept in colloidal suspension, in milk. Thus, caseins are important for the transport of large amounts of calcium to the young, as well as being a vital source of protein in the diet (see Hambraeus, 1982; Schmidt, 1982; Walstra and Jenness, 1984).

A greater diversity of proteins is present in the whey fraction. Some of these are synthesised in the mammary gland, others are derived from blood. The major proteins in ruminants are β-lactoglobulin (BLG) and α-lactalbumin. BLG is a 18,300 dalton protein, present in milk as dimers, interacting non-covalently. Two intracellular disulphide bridges are present. BLG's three dimensional structure has been determined at high resolution (Papiz et al., 1986) and it has been shown to share amino-acid sequence homologies with a number of secretory proteins (see Sawyer, 1987; Ali and Clark, 1988; chapter 6). Its function is unclear, although it is known to be able to bind hydrophobic molecules such as retinol (see chapter 6 for more

details of these topics). On raising the pH from the isoelectric point of BLG (5.1-5.6) to that of milk (6.7), the protein dimerises. Between pH 3.5 and 5.2 it converts to octamers and at pHs below 3.5 reversible dissociation occurs. The dimer also dissociates at alkaline pHs.

α-lactalbumin is a nearly spherical, very compact globular protein, which appears to have much disordered structure, as determined by circular dichroism measurements (see Swaisgood, 1982). It has recently been crystallised and its three dimensional structure has been partially determined and appears to be similar to the structure of lysozyme (Smith et al., 1987). α-lactalbumin contains eight cysteines, forming four disulphide bridges (Vanaman et al., 1970). Its amino-acid sequence shows good homology to hen egg-white lysozyme (Brew et al., 1967). 49 out of 123 amino-acids of bovine α-lactalbumin and hen egg-white lysozyme are identical, with 23 conservative changes (Brew et al., 1970). α-lactalbumin is monomeric in solution and plays an essential role in the biosynthesis of lactose. The protein galactosyl transferase normally transfers galactose from uridine diphosphate to N-acetylglucosamine (in the glycosylation of proteins). α-lactalbumin acts as a specifier protein which binds to galactosyl transferase and causes transfer of galactose to glucose, making lactose.

An iron-binding protein, lactoferrin, is synthesised in the mammary gland (but also in some other secretory organs, although apparently not in the liver - see Walstra and Jenness, 1984). Another iron-binding protein, transferrin, synthesised in the liver and usually present in the blood serum, may be present as well as, or instead of, lactoferrin. Low levels of lactoferrin are found in bovine milk (see tables 1.1 and 1.2). Lactoferrin and transferrin both appear to bind two molecules of trivalent iron per protein molecule. The two proteins show little immunological similarity and amino-acid sequence comparisons show great differences between the

two, although they are clearly related. These proteins have been implicated in bacteriostatic actions by mopping up available iron, thereby depriving iron requiring bacteria of iron.

Blood serum albumin is synthesised in the liver and transported in the blood serum. Blood contains 30-40 g/litre, in milk about 0.4 g/litre is found (Walstra and Jenness, 1984). The exact means of serum albumin transport into milk is not known, the presence of specific receptors in the mammary gland has been postulated (see Jenness, 1982), but it may simply enter milk through intercellular spaces. The function of serum albumin in the milk is unknown.

Immunoglobulins in the milk are derived either from the blood stream, or are synthesised in B cells which are localised in the mammary gland (see Butler, 1974). Immunoglobulin A (IgA) is the principal immunoglobulin in the colostrum and milk of species which transmit immunoglobulin G (IgG) to the young *in utero* (for example, man - see table 1.3). IgA in milk usually occurs as SIgA, dimerised IgA joined by disulphide bridges through a polypeptide, the J-component and by another polypeptide called the secretory component (SC). The SC polypeptide is involved in secretion of IgA into fluids, such as milk (see Butler, 1974; Jenness, 1982). Free secretory component (FSC) is also found in milk. IgG is the principal immunoglobulin in species which transfer little IgG to their offspring i*n utero* (for example, cattle - see table 1.3). The immunoglobulins are involved in transmitting passive immunity to the young prior to its own immune system becoming fully functional.

There are considerable differences between colostrum (the first milk) and "normal milk". In general, the colostrum is much more concentrated than normal milk. Much of the immunoglobulin transfer occurs in the colostrum, immunoglobulin levels falling after the first week post partum. In cows, greater than 50 g/litre of immunoglobulins (most of which is IgG) is present in colostrum. This falls, in normal

milk, to less than 1000 mg/litre (table 1.3). Similarly, lactoferrin levels drop from about 1250 mg/litre to 100 mg/litre within the first week of lactation, in cows.

The figures presented in table 1.2 and 1.3 show the differences in milk protein composition found in different species. The caseins are present in all mammalian milks although only one α_S-like casein is present in many animals. α-lactalbumin is also present in most mammals, except the California sea lion, which contains no lactose in its milk. α-lactalbumin from red-necked wallaby (a marsupial) has been sequenced (Shewale et al., 1984), so its presence in milk occured very early in mammary gland and milk evolution (Hayssen and Blackburn, 1985; Prager and Wilson, 1988). β-lactoglobulin was originally thought to be present only in ruminants, but has now been sequenced in dogs, pigs, horses and in marsupials (see Jenness, 1982; Godovac-Zimmermann, 1988). It is however, absent from human and rodent milks.

In rodents another protein, apparently absent from other classes of mammals, whey acidic protein (WAP), is found. It is a 14 kdal polypeptide with a high cysteine content. Its function in milk is unknown. It appears to share structural similarities with neurophysin, snake venom toxins, wheat germ agglutinin and ragweed pollen allergen Ra5 (Hennighausen and Sippel, 1982). Neurophysins are hypothalamic carrier proteins for oxytocin and vasopressin (see Hennighausen and Sippel, 1982). The proteins share no apparent amino-acid homologies, but have similar positioning of cysteine residues. This has indicated that they may fold to give similar three dimensional structures. The cysteines are similarly positioned in each of these proteins, having a calculated probability of 1 in 10,000 of arising by chance (Drenth et al., 1980) and seems unlikely to have occurred by chance, but may either be due to evolutionary relatedness or convergence of structure. In either case the probable similar three-dimensional folding of these small proteins suggests that they

may have similar roles. A recently described human breast cancer-associated protein, PS2 (see Jeltsch et al., 1987; Rio et al., 1988) shares strong amino-acid homology with a secreted pig protein, porcine spasmolytic protein (PSP) (Rio et al., 1988; and references therein). Inspection of the conserved regions of these two proteins suggests that they may have cysteine patterns similar to those described for the above proteins (my observations). There is evidence that PS2 is normally synthesised in stomach mucosa cells. If PS2, PSP and WAP are indeed structurally similar, it is possible that WAP may be involved in an antispasmolytic action in rodent milk. A similar arrangement of cysteines is apparently present in red sea turtle protease inhibitor (RTPI) (Dandekar et al., 1982). RTPI and WAP show some amino-acid sequence similarities in the region around the cysteines.

1.2 MAMMARY GLAND DEVELOPMENT AND STRUCTURE

Mammary gland development occurs in two stages. The first encompasses fetal, neonate, pre- and post-puberty steps, the second is the lactation stage. The steps in mammary gland development are similar in most, if not all, mammals. The differences are largely due to differences in life cycles, for example, length of pregnancy.

Fetal development involves the formation of a mammary or milk "streak". This is a single-layered ectoderm which appears first on one, then on the other side. The streak extends from the anterior to the posterior limb bud, the length of the streak depending on the number of eventual teats, which varies in different species. The

mammary streak develops into a mammary "line" which gives rise to lens-shaped mammary buds at intervals. This process apparently involves mammary epithelial cell migration. The lens-shaped bud develops into a bulb-shaped structure which pushes down into an underlying "fat pad" structure. In the mouse about 15-20 ducts have formed by birth. In rodents, a nipple also forms by this stage by the invagination of the epidermis. Proliferation of epidermis around the mammary bud leads to nipple formation in man and cattle. In the mouse, male mammary development proceeds along the same path as in females until two-thirds of the way through pregnancy, when androgen production begins. This inhibits further male mammary development. The mammary bud regresses away from the epidermis to form a blind duct. In man, male mammary development proceeds as far as in the female almost until birth (for example, see Hiba et al., 1977).

At birth females have mammary glands consisting of a well-developed teat and some branching duct system, lying in a fat pad. Further development proceeds along with growth of the rest of the body until just before puberty. Some ductal development and branching occurs, keeping pace with general growth.

Just prior to puberty mammary gland development increases. Ductal growth and branching occurs. This growth is dependent on ovarian hormones. The development continues into puberty and reaches a point at which cycles of development occur, during the oestrus cycle. With each cycle more branching and growth occurs. However, this growth is followed by regression in the luteal phase. Nevertheless, the regression is never as great as the proliferation preceding it and a general trend of slow growth is seen.

Far less development occurs during pre- and post-puberty than during pregnancy. During pregnancy the greatest mammary development is achieved. This development is divided into two phases, "mammogenesis" (the growth phase) and

"lactogenesis" (milk secretion).

In the mammogenic phase the mammary gland undergoes further dichotomous branching and replaces the fat pad almost entirely, by parturition. Massive proliferation of the terminal buds occurs, forming the so-called alveoli. This "lobulo-alveolar" structure appears rather like bunches of grapes under the light microscope. They consist of a lumen surrounded by the mammary epithelial cells which are the secretory cells. Behind the layer of epithelial cells are myoepithelial cells whose processes form a sheath around the secretory cell layer. This proliferation sees a vast increase in secretory cell number. In rats and mice a 200-300% increase is seen and in cattle the udder size increases from a few grams to several kilograms. Mammary growth is almost complete by birth in many mammals, including cattle, but continues into early lactation in others, such as rabbits.

The second phase, lactogenesis, begins during mid- to late pregnancy. A secretion appears in alveolar lumens. It is possible that it is at this stage, or just prior to this stage, that induction of milk protein gene expression occurs. In any case, a yellowish fluid first appears in the mammary ducts of many mammals (although not all) at this stage (see chapter 3 for histological and RNA data on lactogenesis in sheep). At parturition a concentrated milk, the colostrum, forms the first milk. This soon gives way to a less concentrated milk differing in some components, as lactation gets underway (see below).

Towards the end of lactation, when the young are no longer entirely dependent on milk for nutrition, there is a gradual loss of cells by cell shrinkage and lysis, or expulsion from the alveolus. The last secretions may contain this cell debris. This process is greatly accelerated after weaning, mammary gland regression continuing until it is back to a state similar to that prior to pregnancy, in readiness for the next pregnancy. For reviews and further references on mammary structure and

development see Hollman (1974), Cowie (1984), Knight (1984), Sakakura (1987), Russo and Russo (1987) and Daniel and Silberstein (1987).

Mammary development occurs in response to several hormones. It is not clear whether there is hormonal requirement for prenatal development, or as to which hormones may be involved. Development of the early embryonal mammary gland is controlled by the mesenchyme surrounding it (Sakakura et al., 1982). Mammary mesenchyme is capable of converting chick epidermis from feather formation to the formation of mammary bud-like structures. Conversely, chick mesoderm can cause mammary bud de-differentiation and feather follicle structure formation (see Cowie, 1984). Local interactions therefore, appear to control mammary gland development in the embryo. A similar experiment was carried out by Sakakura et al. (1976). They transplanted mammary epithelial cells (separated from mammary mesenchyme) from 14-day embryo, combined with salivary mesenchyme, under the kidney capsule of an adult female mouse. A "salivary gland" structure was formed by the mammary epithelial cells. However, during pregnancy and lactation, ducts formed and "milk" was produced. Thus, the mammary epithelium is primed early in its development to lactational potential, although signals from surrounding tissue determine the type of structures formed. These experiments also show that the embryonal mammary epithelium is capable of differentiation to the lactogenic state by lactogenic hormones. This potential has also been shown for post-natal, pre- and post-pubertal mammary epithelial cells in the presence of lactogenic hormones *in vitro* (see Banerjee, 1976; Topper and Freeman, 1980).

At least the later stages of embryonal mammary development may involve the action of hormones. Certainly, androgens halt and cause regression of embryonal mammary development in mouse and rat males (see above). This is a response to androgens and not to the absence of ovarian hormones, as ovariectomy does not cause

this, whereas injection of androgens into females causes male-like mammary gland regression (see Cowie, 1984; Sakakura, 1987). The lactogenic potential of prenatal mammary epithelium is shown by the production of "witches' milk" in newborn babies. This secretion appears to be produced under the influence of maternal or placental hormones and its production stops shortly after birth (Hiba et al., 1977).

Little development occurs until shortly before puberty, when cyclical production of ovarian hormones begins. Prolactin may also be important here since ovarian hormones cause the release of prolactin (see Topper and Freeman, 1980). Serum concentrations of prolactin are greatest during the follicular phase and fall during the luteal phase of the oestrus cycle, in cows and rats. Growth hormone levels increase during pro-estrus in mice (see Tucker, 1974; Anderson, 1974). Increased growth during the follicular phase is followed by a decline (for example, in rats - Sinha and Tucker, 1969a). Similarly, mammary epithelial cells in cows developed just before oestrus and regressed just after oestrus (Sinha and Tucker, 1969b). The regression is never as great as the growth preceding it, so over a number of oestrus cycles ductal growth is seen to have increased.

The massive increase in mammary gland lobulo-alveolar development and mammary gland size during pregnancy requires the presence of several steroid and peptide hormones. These hormones are secreted by the ovaries, anterior pituitary and the placenta. Prolactin is most important during lactation, although corticosteroids, growth hormone and thyroxine are also lactogenic (see Tucker, 1974; Forsyth, 1983; and references therein).

Serum progesterone levels rise throughout pregnancy in sheep, an increase of 10-fold in levels. Around parturition levels drop dramatically to pre-pregnancy levels. Oestrogen levels also fall at parturition. In cows, low prolactin and growth hormone levels increase sharply at parturition (see Tucker, 1974; Heap and Flint,

1984; Cowie, 1984; and references therein).

During pregnancy oestrogen and progesterone act synergistically with the anterior pituitary hormones (prolactin and growth hormone) to give lobulo-alveolar development. Prolactin appears to be particularly important for mammary growth, in sheep (see Forsyth, 1983). The minimum hormonal requirement for rat mammary gland development is insulin, aldosterone, oestrogen, progesterone and prolactin (Wood et al., 1975).

The action of prolactin may be carried out by placental lactogen. The dependence (and the extent of dependence) on either/or, or both hormones varies from species to species. Levels of placental lactogen rise during pregnancy in all mammals. The number of fetuses (and therefore, placenta) can influence the extent of mammary development and milk yield (Knight et al., 1986; and references therein).

Milk secretion begins two-thirds of the way through pregnancy in sheep and many other mammals. In other mammals, such as rabbits, it starts much closer to parturition. Prolactin is the major lactational hormone, but the presence of other hormones can be important for the onset of milk secretion. In the rat, a drop in progesterone levels just before parturition allows placental lactogen, prolactin and corticosteroids to act in lactogenesis (see Cowie, 1984). So, although progesterone is required for mammary growth its presence appears to inhibit lactogenesis. There is also evidence for its inhibition of milk protein gene expression (see below). Thus, progesterone may "hold-off" milk production and secretion until the mammary gland is developed to a certain extent.

In addition to anterior pituitary hormones other signals are also important for lactation. Suckling triggers nerve impulses to the hypothalamus, resulting in oxytocin release into the blood. Oxytocin causes contraction of myoepithelial cells in the mammary gland, causing milk ejection from the alveoli. Prolactin release also

appears to be triggered by suckling, in women (see Grosvenor and Mena, 1974; Tucker, 1974; Forsyth, 1983; Cowie, 1984).

Hormonal control of the involution of the mammary gland at weaning has not been analysed to any great extent. Accumulation of milk that occurs after weaning appears to have an inhibitory effect on further milk synthesis (Russo and Russo, 1987), followed by the involutionary changes described above. In the rat, decreasing milk removal slows prolactin synthesis and release in the pituitary and blocks mammary gland responsiveness to prolactin (see Grosvenor and Mena, 1974), thereby starting a cycle of reduced milk secretion.

Recent work suggests that many growth factors may be involved in mammary gland growth and development during pregnancy. However, their role in these processes is not yet clear (see Dembinski and Shiu (1987) for review).

The hormonal requirements for mammary gland growth and development has been studied in a number of systems. Ovariectomized, adrenalectomized, hypophysectomized animals have been used to determine mammary growth and secretory activity, by hormone replacement. Organ and fragment (explant) culture systems and cell culture systems have been employed for determining the effects of specific hormones on growth and secretion. These *in vitro* approaches are desirable for better control of signals and responses. Cell culture systems, in particular, are important as local tissue-level controls can be removed. The requirements for specific hormones at different stages of mammary gland development have been determined using these systems (see Anderson, 1974; Grosvenor and Mena; 1974; Banerjee, 1976; Topper and Freeman, 1980; Knight and Peaker, 1982; Forsyth, 1983; Knight, 1984; Houdebine et al., 1985; Mepham, 1987; The mammary gland, ed. Neville and Daniel. (1987) Plenum Press).

These studies show that there are many apparent differences between

different species in their hormonal requirements for mammary gland development and lactation. Nevertheless, minimal hormone requirements have been determined for a number of species. Most work has involved mouse and rat systems. In mice, oestrogen, progesterone, prolactin and/or growth hormone are required for *in vivo* mammogenesis. Lactogenesis requires the presence of prolactin (and/or growth hormone) and corticosteroids. Growth hormone and prolactin appear to be equivalent signals in the mouse.

The roles of the individual hormones are not entirely clear. Prolactin, placental lactogen and growth hormone appear to have similar effects on mammogenesis and lactogenesis. It has been suggested (see Topper and Freeman, 1980; Thordarson and Talamantes, 1987) that growth hormone may be more important for mammary gland development during adolescence and placental lactogen dominant in alveolar formation during the second half of pregnancy. Prolactin appears to be the dominant hormone in epithelial growth after parturition. In mouse, serum prolactin levels are high early in pregnancy but have declined by about day nine of pregnancy, when placental lactogen levels in the serum become detectable and increase sharply to a plateau until parturition. Growth hormone levels also increase in the second half of pregnancy but fall before parturition (Thordarson and Talamantes, 1987). The initial high levels of prolactin in the serum have led to suggestions that it may initiate differentiation in early pregnancy (Topper and Freeman, 1980). Furthermore, hypophysectomy during pregnancy does not appear to affect mammary cell number or lobulo-alveolar development during the period when placental lactogen levels in the serum are high. Prolactin appears to be essential for initiation of milk synthesis and important for maintenance of lactation. Hypophysectomy, in ruminants, during lactation leads to decline in milk yields. This decline can be reduced by the presence of growth hormone, glucocorticoids and thyroid hormone. In hypophysectomized

ruminants this treatment gives milk yields 2/3 lower than normal (Cowie and Tindal, 1971).

In vitro experiments (Ceriani, 1970a, b and others - see Banerjee, 1976; Topper and Freeman, 1980; Haslam, 1987 for reviews) suggest that oestrogen is required for duct formation (together with prolactin) during adolescence and for lobulo-alveolar growth during pregnancy. Also, oestrogen levels rise just pre-partum, suggesting a possible role in lactogenesis (see Topper and Freeman, 1980; Haslam, 1987) although results of in vitro experiments are contradictory as to its positive or negative effects on lactogenesis (Kleinberg et al., 1982; Bolander and Topper, 1979). Oestrogen appears to act synergistically with progesterone during pregnancy. Progesterone action appears to involve stimulation of lobulo-alveolar growth. Its presence is not essential in *in vitro* mammogenesis and lactogenesis experiments, using organ culture (Ceriani, 1970a, b). As described above, progesterone inhibits lactogenesis and may do so by acting antagonistically with glucocorticoids (Rosen et al., 1978).

In organ and explant culture the presence of insulin is essential for viability. Its requirement may be due to the absence of insulin-like growth factors (see Dembinski and Shiu, 1987). Some results with *in vitro* organ and explant systems suggest that neither oestrogen nor progesterone are required for alveolar development. However, since the mammary gland is primed *in vivo* for six days prior to the *in vitro* studies the possibility that residual steroids are acting, cannot be ruled out. In addition, peptide hormones from one species are often used for *in vivo* or *in vitro* studies in another species. This is the case for much of the work involving prolactin, growth hormone and placental lactogen. These hormones are members of a family of related proteins. Human growth hormone shares 85% homology with human prolactin at the amino-acid level. A "heterologous" prolactin, for example, could conceivably be

recognised as growth hormone by mammary gland receptors. These caveats must be considered when interpreting results in which exogenous hormones have been applied (see Topper and Freeman, 1980; Vonderhaar, 1987).

Mammary epithelial cell lines have been difficult to establish in the past. Only recently have a number of new approaches been relatively successful. Initially Emerman and Pitelka (1977) (and see Emerman et al., 1977; Yang et al., 1979) successfully cultures mammary epithelial cells on collagen matrix. Others have since then also attempted these methods (for example, Suard et al., 1983), whilst cultures on reconstituted basement membranes have also been attempted (Li et al., 1987). Most schemes for the production of cell lines have selected for the production of one milk protein. It is not clear whether all milk protein genes are being expressed in these lines. Furthermore, morphological alterations in established cell lines have been noted (Bissell, 1981). Transformed cell lines often do not give milk protein gene expression. Ball et al. (1988) selected for COMMA-1D cells in which the endogenous β-casein milk protein gene can be induced by prolactin *in vitro* and cultured this clone in order to analyse prolactin control of expression.

With the relatively recent cloning of many milk protein genes and the recently described technology for making transgenic mice (Palmiter and Brinster, 1986), *in vivo* analysis of introduced milk protein gene regulation has become feasible (Simons et al., 1987; Lee et al., 1988).

Many assays have been used for analysing the effects of hormones on mammary growth. *In vivo* studies and organ culture allow histological examination of differentiation and growth. For these and for explants and cell culture systems other methods include measurement of changes in DNA and/or RNA content, protein levels, enzyme activities (and RNA, DNA polymerase levels and steroid and peptide hormone receptor levels) and polysome and cytomembrane content (see Banerjee, 1976 for

review of these procedures).

Because of the abundance of milk proteins in the mammary gland, their presence has been used to assay the onset of lactogenesis (see Banerjee, 1976; Topper and Freeman, 1980). The changes in levels of mRNAs encoding caseins and α-lactalbumin have been described (Rosen et al., 1975; Shuster et al., 1976; Nakhasi and Qasba, 1979; Burditt et al., 1981; Hobbs et al., 1982), as have mRNA activities, by presence in polysomes (Gaye et al., 1972), translatability by *in vitro* translation systems (Rosen et al., 1975; Nakhasi and Qasba, 1979) and protein levels in the mammary gland (Shuster et al., 1976; Nakhasi and Qasba, 1979; Burditt et al., 1979). These studies have shown that mRNA and protein levels increase during pregnancy and lactation. More recently the cloning of individual casein cDNAs has allowed differential expression of the different caseins and whey proteins to be determined (Nakhasi and Qasba, 1979; Hobbs et al., 1982; Vonderhaar and Nakhasi, 1986; Pittius et al., 1988). These studies have shown that milk protein genes are non-coordinately regulated.

Many studies have also demonstrated that milk protein genes are hormonally regulated (Mehta et al., 1980; Ganguly et al., 1980; 1982; Shuster et al., 1976; Teyssot and Houdebine, 1980; 1981. See also Banerjee (1976), Topper and Freeman (1980), Houdebine et al. (1985), for reviews). In particular, these studies have suggested that milk protein gene expression is directly affected by the lactogenic hormones (prolactin and glucocorticoids), as addition of these hormones gives rapid changes in mRNA levels (in *in vivo* and *in vitro* experiments). Prolactin stimulates transcription of rat (Guyette et al., 1979) and rabbit (Teyssot and Houdebine, 1980) casein genes. Furthermore, glucocorticoids also stimulate transcription of mouse (Ganguly et al., 1979) and rabbit (Teyssot and Houdebine, 1981) casein genes, in the presence of prolactin. The increase in transcription is apparently not sufficient to

explain the increased casein mRNA levels and it has been suggested that post-transcriptional events increase mRNA stability (Guyette et al., 1979; Teyssot and Houdebine, 1980). Their results indicate that prolactin may primarily influence mRNA stability, whilst glucocorticoids amplify transcriptional activation by prolactin, but do not enhance mRNA stability (Teyssot and Houdebine, 1981). Progesterone appears to inhibit the stimulatory effects of prolactin on gene expression and mRNA stability (Teyssot and Houdebine, 1981). It has also been suggested that progesterone acts through antagonism with glucocorticoid action and progesterone competition of dexamethasone binding to cytoplasmic receptors, has been reported (Ganguly et al., 1982).

Much recent work on transcriptional activation by DNA-binding proteins has suggested many mechanisms by which transcription may be modulated (see chapter 4 for references and discussion). Glucocorticoids and progesterone bind to their receptors, DNA-binding proteins which directly regulate transcription of target genes. Prolactin, on the other hand, binds to a cell surface receptor and may modulate transcription through the action of secondary messengers (although no such messengers have been found - see Boutin et al., 1988). The synergistic actions of transcriptional activators have been well characterised. Antagonistic actions of glucocorticoid and progesterone receptors have been suggested by the finding that glucocorticoid and progesterone receptors bind to similar, if not identical, DNA recognition sequences (see chapter 4).

Many workers have reported differential expression of milk protein genes (see above). Variable promoter strengths and mRNA stabilities may account for these differences. In addition, work by Ono and Oka (1980a, b) shows that different concentrations of cortisol elicit maximal α-lactalbumin and casein synthesis. They found that cortisol concentrations of 10^{-8} M gave maximal α-lactalbumin synthesis,

but little casein synthesis, in virgin mouse mammary gland explant culture. Casein synthesis required 10^{-6} M cortisol, a concentration at which little α-lactalbumin was synthesised (i.e. this concentration was inhibitory for α-lactalbumin synthesis). Changes in glucocorticoid levels occur during pregnancy, levels rising during the second half of pregnancy to about 10^{-6} M and falling after parturition to 10^{-8} M.

1.3 MILK PROTEIN GENE STRUCTURE AND EXPRESSION

cDNAs for most or all guinea pig, mouse, rat, bovine, ovine and rabbit caseins have been cloned (see Bonsing and Mackinlay, 1987; Mercier et al., 1985; Devinoy et al., 1988) and a cDNA for human β-casein has also recently been sequenced (Menon and Ham, 1988). In addition, the gene structures of rat α-, β- and γ-caseins (rat α-casein is orthologous with bovine α_{s1}-casein and rat γ-casein is orthologous with bovine α_{s2}-casein) (Yu-Lee et al., 1986; Jones et al., 1985; Yu-Lee and Rosen, 1983), bovine α_{s1}-casein (Yu-Lee et al., 1986) and bovine β-casein (Gorodetsky et al., 1988) have been determined. Partial characterisation of rabbit α_{s1}- and β-casein genes have been described (Devinoy et al., 1988). Comparison of the calcium-sensitive, α_{s1}-, α_{s2}- and β-caseins show that they are evolutionarily related.

The caseins are highly diverged from each other, with low amino-acid sequence homologies. Three main regions of conservation have been described. The leader, or signal, peptide is fifteen amino-acids in length in all sequenced calcium-sensitive caseins. The nucleotide sequences around the major phosphorylation

sites are also highly conserved. Finally, alignment of the 5' untranslated sequences shows conservation, suggesting that the 5' untranslated sequences may be involved in the formation of secondary structure important for expression (see Bonsing and Mackinlay, 1987).

Most work on the gene structures of caseins has been carried out on the rat genes by Rosen's group. They range in size from 7.5 kb (β-casein) to 15 kb (γ-casein). α-casein is probably 10-15 kb long (Yu-Lee et al., 1986). Rat β-casein gene contains nine exons (as probably does bovine the β-casein gene - see Bonsing and Mackinlay, 1987); the first and last exons are entirely non-coding (Jones et al., 1985). The rat γ-casein is also encoded by a nine exon gene (Yu-Lee and Rosen, 1983). Comparison of amino-acid and cDNA sequences and of genomic organisation has suggested that the ancestral casein gene underwent internal duplication of one exon to four exons, exons III-VI, and that this gene duplicated to give the α_S- and β-casein genes. In addition to the conserved transcribed regions, the 5' flanking sequences of the rat α-, β- and γ- and bovine α_{S1}-casein genes show some conservation. These conserved sequences may be important in the mammary specificity and hormonal regulation of casein gene expression (see reviews by Stewart et al., 1987; Bonsing and Mackinlay, 1987; Rosen, 1987).

κ-casein cDNAs have been cloned from cow, mouse, rat and sheep (Stewart et al., 1984; Kang and Richardson, 1988; Thompson et al., 1985; Mercier et al., 1985; J-L. Vilotte, personal communication). Bovine κ-casein gene is at least 10 kb long and has not yet been sequenced (see Bonsing and Mackinlay (1987) for citation of unpublished results), so it is not possible to determine whether the κ-casein and γ-fibrinogen gene structures are similar. Gene structure comparison would test the hypothesis of their relationship. Comparison of protein and cDNA sequences strongly suggests that the calcium-insensitive κ-casein is not related to the other caseins.

Gupta et al. (1982) showed that mouse α-, β- and γ-caseins map to mouse chromosome 5, using somatic cell hybrids. Similarly, Gaye et al. (1986) have mapped the sheep α_{s1}-, α_{s2}-, β- and κ-caseins to the same arm of sheep chromosome 2. Linkage of the four bovine caseins has also been reported by Matyukov and Urnyshev (1980). This is surprising if κ-casein is indeed unrelated to the other casein genes.

The rat, human, bovine and guinea pig α-lactalbumin genes have been cloned (Qasba and Safaya, 1984; Hall et al., 1986; Vilotte et al., 1987; Laird et al., 1988). The genes are well conserved and contain four exons. Qasba and Safaya (1984) compared the rat α-lactalbumin gene with the chicken lysozyme gene. The chicken lysozyme gene is also encoded by four exons. Exons I-III are similar in size and encode similar numbers of amino-acids). Exon IV of the α-lactalbumin gene, however, is considerably larger than exon IV of the lysozyme gene and encodes extra amino-acids. Sequence similarities of 43-56% were noted.

The major whey protein of rodents, WAP, is encoded by a 2.8 kb, four exon gene (Campbell et al., 1984). Partial characterisation of the rabbit WAP gene has been presented by Devinoy et al. (1988).

The ovine BLG gene has been cloned (A. J. Clark's results; Ali and Clark, 1988). It is encoded by a 4.9 kb, seven exon gene (presented in this thesis; see chapters 4 and 5), which shows a similar gene organisation to the serum retinol binding protein, rodent major urinary protein, apolipoprotein-D and α1-acid glycoprotein genes (see chapter 6). The bovine BLG gene has been cloned (unpublished results cited in the review by Bonsing and Mackinlay, 1987) and it appears to be very similar to the ovine gene (A. G. Mackinlay, personal communication).

1.4 AIMS OF THIS PROJECT

Gene cloning has shown that milk proteins are encoded by abundantly expressed, single copy genes. They should, therefore, be very useful for studying hormonal control of mammary gland development. The mechanisms controlling milk protein gene induction and expression are far from clear. The expression of these genes increases greatly during pregnancy and lactation, due to transcriptional and/or post-transcriptional processes. Glucocorticoids and prolactin activate and sustain milk protein gene expression.

When this project was initiated little more than the above was known about the transcriptional control of milk protein genes, that is, the DNA sequences which are important for tissue-specific and temporal regulation of milk protein gene expression. The recently established technology for making transgenic mice (Palmiter and Brinster, 1986) suggested an alternative method for studying transcriptional control of these genes in the absence of suitable cell culture systems. Furthermore, a project to target foreign gene expression to the mammary gland, for the production of medically important human proteins in milk, was initiated (Lathe et al., 1986). Lathe et al. (1986) proposed that transgenic sheep (or cattle) could be milked to obtain large quantities of these proteins.

The gene encoding ovine BLG was cloned to provide sequences for targeting human gene expression to the mammary gland in transgenic animals. Preliminary characterisation and DNA sequencing were initiated by A. J. Clark (Ali and Clark, 1988). Mice transgenic for the BLG gene were shown to give abundant expression of the gene in the mammary gland (Simons et al., 1987). Subsequently, Lee et al. (1988) have demonstrated tissue-specific expression of the rat β-casein gene at low levels in transgenic mice. Work by S. Harris (unpublished results) suggests that the ovine BLG gene is appropriately regulated in transgenic mice. BLG gene expression, in transgenic mice, follows a similar time course of expression as that in sheep

(described in chapter 3). The 16.2 kb region containing 4.0 kb of 5' flanking sequences, 7.3 kb of 3' flanking sequences and the 4.9 kb transcription unit therefore contains sequences sufficient for "correct" expression of the ovine BLG gene.

Production of transgenic sheep carrying human liver-specific genes (cDNAs for human Factor IX and α_1-antitrypsin) has been demonstrated (Simons et al., 1988) and low levels of expression observed (Clark et al., manuscript submitted; Archibald et al., unpublished results). Transgenic mice carrying the human α_1-antitrypsin gene, driven by ovine BLG 5' flanking sequences and 5' untranslated sequences, make moderately high levels of human α_1-antitrypsin in milk (Archibald et al., unpublished results). Low levels of a human tissue plasminogen activator cDNA, fused to the 5' flanking sequences of the mouse WAP gene, have been demonstrated (Gordon et al., 1987). WAP-ras (Andres et al., 1987) and WAP-myc (Schonenberger et al., 1988) were expressed tissue-specifically and under hormonal control (during the first lactation). Low levels of expression (50-fold and 10-fold lower than endogenous WAP, respectively) were oberved. These experiments suggest that 5' flanking sequences of the ovine BLG and mouse WAP genes contain sufficient sequences for mammary-specific expression of these genes.

Comparisons of the 5' flanking sequences of milk protein gene structures have indicated the presence of few obvious progesterone or glucocorticoid responsive elements. The significance of the putative responsive elements which have been described has not been tested (see chapter 4 for sequence analysis of the ovine BLG gene sequences). Rosen et al. (1985) described the presence of two putative glucocorticoid responsive elements in the γ-casein gene 5' flanking sequences and a putative progesterone responsive element in the fourth intron of the rat β-casein gene. They also stated that the chicken progesterone receptor can bind to rat γ-casein gene 5' sequences, *in vitro*. However, the functional significance of any of these sequences has

not been demonstrated. A number of regions of sequences similarity between rat α-, β-, γ- and bovine α_{s1}-casein genes have been noted (Yu-Lee et al., 1986). Some of these sequences are also conserved in the rabbit casein genes (Devinoy et al., 1988). Furthermore, Hall et al. (1987) showed that a region of about 30 bp is present at -140 to -110 (relative to the transcriptional start site) in many casein and α-lactalbumin genes from cow, guinea pig, man and rat (see also Laird et al., 1988). Devinoy et al. (1988) describe such a sequence at -110 to -139 in the rabbit β-casein gene. Comparison of the bovine α-lactalbumin gene (Vilotte et al., 1987) for the presence of this sequence suggests that only the 5' portion is present (J-L. Vilotte, personal communication). In all these genes the Hall et al. (1987) sequence motif is present at the same region, relative to the transcriptional start site. The motif appears to be absent from the mouse, rabbit and rat WAP (Campbell et al., 1984; Devinoy et al., 1988) and ovine BLG (chapter 4) genes. Devinoy et al. (1988) show that a sequence similar to the 5' portion of this motif is present at -200 of the rabbit WAP gene. Hall et al. (1987) suggested that this sequence might contain some of the information for tissue-specific expression of milk protein gene expression, because of its presence in so many milk protein genes.

Although sequences which may be important for hormonal control of milk protein gene expression have not yet been determined, Lubon and Hennighausen (1987, 1988) have described work characterising the proximal promoter sequences of the mouse WAP and rat α-lactalbumin genes. Their results with 350 bp of proximal mouse WAP gene 5' flanking sequences (Lubon and Hennighausen, 1987) indicated the formation of four DNA-protein complexes (one of which appears to be the TATA-box-binding factor). They found little evidence for mammary-specific complex formation. Their work describing complex formation at the rat α-lactalbumin gene (Lubon and Hennighausen, 1988) provided no evidence for the presence of

stage-specific nuclear factors, that is nuclear factors present or absent in virgin, pregnant or lactating rat mammary tissue. They also investigated complex formation at the Hall et al. (1987) consensus sequence of rat α-lactalbumin gene and found binding to the 5' part of this sequence. The binding site is similar to a NFI recognition sequence and indeed they found that purified HeLa cell NFI binds to this sequence. Their work, therefore, offers few indications of steroid receptor, tissue-specific or stage-specific DNA-protein complex formation, *in vitro*.

As stated above, the transcriptional significance of none of the milk protein gene DNA sequences has been demonstrated. No successful cell transfection experiments have been reported in mammary- or non-mammary-derived cell lines, although Rosen and coworkers (1985; Bisbee and Rosen, 1987) and Kawamura et al. (1987) have described some preliminary experiments. Work with transgenic animals, described above, has progressed quickly and may yield information about transcriptionally important sequences. Deletional analysis (S. Harris, unpublished results) indicate that 800 bp of 5' flanking, the 4.9 kb transcription unit and 1.8 kb of 3' flanking DNA is sufficient for tissue-specific expression, in transgenic mice (this entire region has been sequenced and is presented in chapter 4). Preliminary results suggest that this construct is temporally regulated in a similar manner to the 16.2 kb construct (S. Harris, unpublished observations). Archibald et al. (unpublished results), in this lab, have shown that 3.9 kb of 5' flanking sequences and +30 bp of exon I sequences can drive efficient mammary-specific expression of the human liver-specific α_1-antitrypsin gene, in transgenic mice. These two results suggest that 800 bp of 5' flanking sequences of the ovine BLG gene are sufficient for efficient, mammary-specific expression in transgenic mice. Deletion constructs of the BLG gene, containing less 5' flanking sequence, have been injected into mice and transgenic mice are currently being analysed for expression. Furthermore, constructs

containing the BLG gene without some, or all, introns are also being tested in transgenic mice (A. J. Clark et al., unpublished data). Therefore, transgenic mouse work may soon map regions required for temporal and/or tissue-specific regulation of the ovine BLG gene.

My project has involved the detailed characterisation of the ovine BLG gene, DNA sequencing to determine its exonic organisation and comparison with the gene structures of evolutionally related. The relationship between two similar, but differing (in their restriction maps), phage clone types, has also been investigated. Furthermore, the tissue-specificity of BLG and other ovine milk protein genes has been demonstrated. Patterns of expression of the BLG gene during pregnancy and lactation, were analysed and differential expression of the ovine milk protein genes was demonstrated.

Chapter 2 MATERIALS AND METHODS

2.1 SHEEP USED

First pregnancy (gimmer) ewes and fourth pregnancy (draft) ewes were sacrificed at various stages of pregnancy, for analysis of milk protein gene expression (chapter 3). Blackface sheep were used for the time course study. 20-30 mls of blood was taken into evacuated tubes about an hour before the animal was killed. The blood was allowed to clot overnight, at room temperature. Blood serum was then removed into new tubes and stored at -20°C until hormone assays were performed. Serum progesterone and prolactin levels were kindly determined by Dr. A. S. McNeilly (MRC Reproductive Biology Unit, Edinburgh) using the methods described by McNeilly and Andrews (1974) and McNeilly (1984).

One mammary gland and a portion of the liver from each ewe was removed immediately after the ewe had been killed. Approximately 0.5 g of tissue was homogenised, either in guanidinium thiocyanate (for making RNA - see section 2.8.5) and in TE (for making DNA - see section 2.8.4). The remainder of the mammary gland and liver was cut into a number of portions, immediately frozen in liquid nitrogen and stored under liquid nitrogen.

For preparation of blood lymphocyte DNA (chapters 4 and 5), approximately 7 mls of blood was taken into evacuated tubes containing an anticoagulant (heparin). The tubes were kept on ice until processed (see section 2.8.4).

Sheep milk samples were obtained by hand milking. 10-20 mls of milk was taken from each animal and processed (as described in section 2.13.1).

The sheep used in these studies were kept at the Large Animal Unit, Edinburgh (Roslin), IAPGR and at the institute's farm at Blythebank (Peebleshire).

2.2 RECOMBINANT PLASMIDS AND BACTERIOPHAGE LAMBDA.

Ovine milk protein cDNAs, p931 (BLG cDNA), αlac-1030 (α-lactalbumin cDNA), pαs1214 (α_{s1}-casein cDNA), pUC40 (α_{s2}-casein cDNA), pβcas (β-casein cDNA) and pκ551 (κ-casein cDNA) were the kind gifts of Dr. J.-C. Mercier (INRA, France) (Mercier et al., 1985). These cDNAs are all cloned in pBR322, with the exception of pUC40, which has been subcloned into pUC19. The bovine MHC Class I cDNA (named pBoLaI - subcloned into ptg1(poly)) was kindly provided by P. Brown (IAPGR) (Brown et al., 1989). The β-actin cDNA (Ponte et al., 1983) and the Xenopus rDNA probe (Sollner-Webb and Reeder, 1979) were provided by J. O. Bishop.

The bacteriophage recombinants SS1 and SS12 were isolated by Dr. A. J. Clark and encode the ovine BLG gene. The bacteriophage vector used was EMBL3 (see Ali and Clark, 1988).

Subclones of the bacteriophage SS1, SS1BH, SS1HX and SS1 SalI/SphI have been constructed in ptg1(poly) (Lathe et al., 1987) or pUC18 (Yanisch-Perron et al., 1985) by Dr. A. J. Clark and Dr. S. Harris.

All other recombinants were my own constructs and are described elsewhere in the thesis.

2.3 NONRECOMBINANT PLASMIDS, BACTERIOPHAGE AND BACTERIAL STRAINS

The pBR322-derived vector plasmid, ptg1(poly) (Lathe et al., 1987) and pUC18 (Yanisch-Perron et al., 1985) were used to subclone from the recombinant lambda clones, SS1 and SS12. The M13 bacteriophage vectors, tg130 and tg131 (Kieny et al., 1983) were used in cloning of DNAs for M13 sequencing (see section 2.11).

The *Escherichia coli* bacterial strains listed below were used for preparing plasmid and M13 DNA :

JM83 (*ara*, Δ(*lac-pro*), *strA*, *thi1*, *Φ80dlacI*^{q}, ZΔM15)

JM101 (Δ(*lac-pro*), *supE44*, *thi1*, F'*traD36*, *proAB*, *lacI*^{q}, ZΔM15)

HB101 (F^{-}, *hsdS20* (r_B^-, m_B^-), *recA13*, *ara-14*, *proA2*, *lacY1*, *galK2*, *rpsL20* (Sm^{r}), *xyl-5*, *mtl-1*, *supE44*, λ^{-})

DL43 (F^{-}, *dam*$^{-3}$, *dcm*$^{-6}$, *MetB1*, *galK2*, *galT22*, *lacY1*, *tsx-78*, *supE44*, *thi-1*, *tonA31*, *mtl-1*)

JM83 (Yannisch-Perron et al., 1985), JM101 (Yannisch-Perron et al., 1985) and HB101 (Bolivar and Backman, 1979) were obtained from Dr. J. O. Bishop (Dept. of Genetics, Univ of Edinburgh) and DL43 was obtained from Dr. D. Leach (Dept. of Mol. Biol., Univ. of Edinburgh).

2.4 ENZYMES, ANTIBIOTICS, CHEMICALS AND REAGENTS

Unless otherwise stated, these were all purchased from the companies listed below.

Amersham International Plc, Amersham UK, Aylesbury, England.

BCL Boehringer Mannheim, Lewes, England.

BDH Ltd., Glasgow, Scotland.

Becton Dickinson UK Ltd., Plymouth, England.

Fluka Chemie AG, CH-9470 Buchs, Switzerland.

Gibco BRL, Gibco Ltd., Paisley, Scotland.

New England Biolabs, Beverly, MA, USA.

Pharmacia Ltd., Milton Keynes, England.

Sartorius Ltd, Surrey, England.

Serva Feinbiochemica, Heidelberg/New York.

Sigma Chemical company Ltd., Poole, England.

Sterilin Ltd., Feltham, England.

2.5 AUTORADIOGRAPHY AND PHOTOGRAPHY

Hybridisation filters and DNA sequencing gels were autoradiographed using AGFA CURIX X-ray film. All DNA sequencing gels were autoradiographed at room temperature without any intensifying screens. Unless otherwise stated, all hybridisation filters were autoradiographed in the presence of two intensifying screens, as described by Maniatis et al. (1982), for varying periods of time. For these filters the autoradiographic film was pre-flashed, as described by Maniatis et al. (1982) and the film was exposed at -70°C.

Autoradiographic X-ray films were developed by immersion in developer (Kodak LX 24) for 5 minutes, followed by a rinse in water, and 5 minute immersion in fixer (Kodak FX-40) containing a hardener (Kodak HX-40). The autoradiographs were then immersed in circulating tap water for 10-20 minutes and dried at room temperature or in an 80°C oven.

DNA and RNA in agarose gels were visualised by ethidium bromide staining, using a short wavelength ultraviolet trans-illuminator (Ultraviolet Products Inc., Cambridge, England). The stained gels were photographed using a Polaroid MP4 camera and Polaroid 667 Instant Film using variable exposure times.

2.6 MEDIA AND GENERAL DNA AND RNA HANDLING TECHNIQUES

2.6.1 Media

The liquid and solid media described below were used. See sections 2.6.2, 2.8.1 and 2.8.2 for applications.

Luria Broth (LB) - 10 g bactotryptone, 5 g bacto yeast extract, 10 g NaCl, per litre of distilled water. Adjust to pH 7.5 with NaOH. In addition, 15 g of bacto-agar was added, for making plates; or 6 g of bacto-agar was added to prepare soft agar overlays.

BBL agar - 10 g BBL trypticase, 5 g NaCl, 10 g agar (or 6 g agar for soft overlays), per litre of distilled water.

Minimal medium plates - 7.5 g agar, 100 ml 5 x M9 salts (35 g Na_2HPO_4,

15 g KH_2PO_4, 2.5 g NaCl, 5 g NH_4Cl, per litre), 400 ml H_2O, 0.5 ml 1 M $MgSO_4$, 0.5 ml 0.1 M $CaCl_2$, 0.5 ml thiamine hydrochloride (0.2 mg/ml), 6 ml 20% (w/v) glucose. These were used to maintain the JM101 phenotype. This medium selects for maintenance of the F', since these cells require functions present on the F' plasmid for viability in minimal medium. F' functions are required for M13 entry into *E. coli* cells (see Yannisch-Perron et al., 1985).

2.6.2 DNA/RNA

Phenol/chloroform extractions

0.5 volume of phenol (equilibrated with H_2O) was mixed with the DNA solution and briefly centrifuged to separate the phases. The upper, aqueous phase containing the DNA, was removed to a new tube. 0.5 volume of phenol and 0.5 volume of chloroform/isoamyl alcohol (24:1) (H_2O saturated) were added, mixed and the upper aqueous phase was removed into another tube. 0.5 volume of chloroform/isoamyl alcohol was added, mixed and centrifuged. The upper phase was removed to the final tube and the DNA was ethanol precipitated. The phenol used had been distilled over zinc, equilibrated with H_2O and stored at 4°C.

Ethanol precipitation

DNA was precipitated from solution by the addition of 0.1 volume of 3M sodium acetate, pH 5 and 2-3 volumes of ethanol. After mixing, the tube was placed at -70°C for 30-60 minutes or at -20°C, overnight. Precipitated DNA was pelleted by centrifugation at 10,000 rpm for 20 minutes, at 4°C (or in a bench eppendorf centrifuge by spinning for 10 minutes, at maximum speed). The supernatant was removed and the pellet was washed with 70% ethanol (70% ethanol, 30% TE),

centrifuged briefly and the 70% ethanol was removed. DNA pellets were vacuum-dried and resuspended in the appropriate volume.

This protocol was also followed for ethanol precipitation of RNA, except where stated otherwise.

Restriction endonuclease digestion

DNA was digested by restriction enzymes, in buffers recommended by the appropriate manufacturer, at the recommended temperature. Digestions were carried out in volumes containing $\leq$ 0.2 µg/µl DNA. Plasmid and bacteriophage DNA digestions were normally carried out for 1-2 hours, whereas genomic DNAs were often digested overnight. Digests were checked for completion by running on small agarose gels and enzymes inactivated by heating at 70°C for 15 minutes and/or by adding Na_2EDTA to a final concentration of 20 mM. Whenever appropriate, the DNA was phenol extracted and ethanol precipitated.

RNase A (10 mg/ml in 2 x SSC, preboiled for 10 minutes and stored at -20°C) was sometimes added to the restriction digest to a final concentration of 30 µg/ml.

DNA ligation

Restriction enzyme digested DNAs were ligated in a total volume of 20 µl, at 14°C. Ligation mixes were generally incubated overnight, but were sometimes incubated for 2 hours, when ligating "sticky-ends" or 4 hours when ligating blunt ended restriction enzyme cuts. No more than 0.2 µg of total DNA was ligated in such a mix. The reaction mix was prepared by mixing DNA A + DNA B (to a final volume $\leq$ 10 µl), 2 µl 10 x Ligase buffer (0.66 M Tris-HCl pH 7.6, 10 mM Na_2EDTA pH7.0,

0.1 M $MgCl_2$, 0.4 M NaCl), 2 µl DTT (100 mM), 2 µl rATP (2.5 mM), dH_2O, to a final volume of 20 µl and 1 unit of T4 DNA ligase.

Making Competent Cells

Competent cells were prepared by the method of Hanahan (1983). The *E. coli* strains were grown in LB medium, with gentle shaking, overnight at 37°C. The culture was diluted 1/100 in LB medium and grown at 37°C, with shaking, to an OD_{540} of 0.4-0.6. The culture was centrifuged at 7000 rpm for 5 minutes. The supernatant was discarded, the cell pellet resuspended in 1/5 volume with ice-cold CM1 buffer (10 mM Na-Acetate pH 5.6, 5 mM NaCl, 50 mM $MnCl_2$) and placed on ice for 20 minutes. This was spun at 7000 rpm for 5 minutes and resuspended in 1/50 original volume ice-cold CM2 buffer (10 mM Na-acetate, 5% glycerol, 70 mM $CaCl_2$, 5 mM $MnCl_2$). 100 µl of these "competent" cells were used for each transformation reaction. The competent cells can be stored at -70°C for short periods of time.

Transformation with plasmid DNA

Less than 100 ng of DNA was used for each transformation. The DNA was added to 100 µl of competent cells and placed on ice for 30 minutes. The cells were heat shocked at 42°C for 2 minutes, 1 ml LB medium was added and the cells were placed at 37°C for 60 minutes to allow recovery, then plated on LB agar plates (+ antibiotic) and incubated overnight at 37°C.

Transformation with M13 DNA

100 µl of JM101 competent cells were used per transformation. Less than 100 ng of total DNA was added and incubated on ice for 40 minutes. The cells were heat shocked at 42°C for 2 minutes. 100 µl of cells grown to the same density as the competent cells (and kept on ice) were added to the transformed cells, as lawn cells. 3 ml BBL overlay agar (containing 12 µg/ml X-gal, 6 µg/ml IPTG), cooled to 45°C, was added and the mixture was poured onto BBL agar plates and allowed to set. The plates were incubated at 37°C, overnight. White plaques are indicative of recombinant clones (due to disruption of the *lacZ* gene present in the vector) and were picked and DNA was prepared, as described in section 2.8.3.

2.7 GEL ELECTROPHORESIS

2.7.1 DNA gels

Horizontal, buffer immersed agarose gels were run in 40 mM Tris-acetate (0.04 M Tris-acetate, 0.002 M Na_2EDTA), 2 mM Na_2EDTA, pH 7.7 buffer containing ethidium bromide at 0.5-1.0 µg/ml. Two sizes of gels were run. Minigels (10 x 10 cm) were used for quick analysis for completion of digestion, etc. and 20 x 20 cm gels were used for restriction mapping and Southern transfer. Agarose concentration varied between 0.4 - 1.5% (w/v), depending on the size of fragments to be separated, according to Maniatis et al. (1982). Gel combs used depended on the volume of sample to be loaded. Before loading, 0.1 volume of ficoll marker dye (30% ficoll, 0.1% SDS, 40 mM Na_2EDTA, 1.2 mg/ml bromophenol blue) was added to each sample. The samples were then heated at 70°C for 10 minutes and loaded. Gels were

run at varying voltages but restriction mapping gels and gels to be transferred were run overnight at 25-35 volts, until the marker dye had run 2/3 - 3/4 of the length of the gel.

DNA was visualised using an ultraviolet trans-illuminator (section 2.5).

2.7.2 RNA gels.

RNA samples were run on formaldehyde-MOPS agarose gels. 400 ml 1.5% agarose gels were used. 6 g of agarose was added to 280 ml dH_2O. The agarose was dissolved by heating in a microwave and allowed to cool to about 60°C. 40 ml 10 x MOPS (0.2 M MOPS, 0.05 M Na-acetate pH 7.0, 0.01 M Na_2EDTA; pH 7.0 with NaOH) and 80 ml formaldehyde were then added and a 20 x 20 cm gel was poured.

The RNA samples were denatured before loading. The RNA was present in a volume of $\leq$ 15 µl (ethanol precipitated and resuspended in this volume, if necessary). 31 µl of formamide (3 x recrystallised, deionised), 6.25 µl 10 x MOPS and 10 µl formaldehyde were added. The sample was denatured at 65°C for 15 minutes, 6.25 µl of marker dye (50% glycerol, 1 mM Na_2EDTA, 0.4% bromophenol blue, 0.4% xylene cyanol) was added and the sample was loaded onto the gel.

The running buffer was 1 x MOPS. The gels were run at 25-35 volts, overnight. The gels were stained with ethidium bromide, if required, after running.

2.7.3 Acrylamide gels

Acrylamide gels were used for DNA sequencing and for S1 analysis. Their preparation is described in section 2.11.1.

2.7.4 Isolation of DNA fragments from agarose gels

A trough was cut in front of the fragment to be isolated. Dialysis tubing (preboiled for 10 minutes in H_2O) was placed in the trough and the DNA was electrophoresed into the tubing. Once the DNA had run in, the tubing was removed and placed in 5 ml of low salt buffer (0.2 M NaCl, 20 mM Tris-HCl pH 7.4, 1 mM Na_2EDTA) and shaken for 20 minutes to get DNA out into the buffer. Schleicher & Schull Elutip d-columns (NA 010/1) were used to clean up the DNA. The column was washed with 5 ml of high salt buffer (1 M NaCl, 20 mM Tris-HCl pH 7.4, 1 mM Na_2EDTA), followed by 5 ml of low salt buffer to equilibrate the column. The DNA solution was then passed through. The column was washed with 5 ml of low salt buffer and then the DNA was eluted from the elutip with 1 ml of high salt buffer, into a siliconised corex tube. H_2O was added to reduce the NaCl concentration to 0.3 M, 15 µl of dextran sulphate (1 mg/ml) and 2-3 volumes of ethanol were added. The ethanol precipitation procedure described above was followed.

Isolation of DNA fragments from acrylamide gels is described in section 2.12.

2.7.5 Slot Blotting

RNA samples were ethanol precipitated, resuspended at 50 µg/ml and [illegible] absorbance at 260 and 280 nm taken to confirm concentrations. 50 µl of each RNA sample (at 50 µg/ml) was placed in an eppendorf tube and 45 µl of formaldehyde and 405 µl dye/dH_2O (0.4% bromophenol blue) were added. The samples were placed at 65°C for 15 minutes, cooled on ice and centrifuged for 5 seconds in a microfuge. 50 µl was applied to each slot (50 µl contained 0.25 µg of RNA).

Before application of the RNA samples, Amersham Hybond nylon membrane

was cut to the size of the slot blot manifold apparatus, rinsed in dH_2O and placed on the apparatus, taking care to allow no air bubbles to become trapped. The apparatus was connected to a water pump for suction of applied liquid through the apparatus. After application of all samples, the filter was removed, rinsed in 50 mM NaPi pH 7.2 and the DNA was fixed to the filter by ultraviolet irradiation in the usual manner.

Hybridisations were carried out as described in section 2.10 and in chapter 3.

2.8 PREPARATION OF DNA AND RNA

2.8.1 Rapid method for small scale plasmid and M13 DNA preparation

This protocol has been adapted from Birnboim and Doly (1979). 3 ml of LB medium (+ antibiotic) was innoculated with a single colony. This was shaken overnight at 37°C. For M13, a single plaque was placed in 3 ml LB medium using a toothpick, a drop of saturated JM101 cells (grown overnight in LB medium) was added. This was shaken overnight at 37°C.

1.5 ml of the overnight culture was poured into eppendorf tubes and centrifuged for 1 minute. The supernatant was discarded, the pellet was resuspended into 100 µl of 2 mg/ml lysozyme (made up in ice-cold 50 mM glucose, 25 mM Tris-HCl pH 8.0, 10 mM Na_2EDTA) and incubated on ice for 30 minutes. 200 µl of 0.22 M NaOH/10% SDS (9 : 1) was added and mixed. After a further 5 minutes on ice, 150 µl of 3 M Na-Acetate (pH 5.0) was mixed in and incubated for 60 minutes

on ice. 380 µl of the clear supernatant was removed to a fresh tube after centrifugation for 10 minutes, ethanol precipitated and resuspended in 250 µl TE (10 mM Tris-HCl pH 7.5, 1 mM Na_2EDTA). This was again ethanol precipitated, as described in section 2.6.2 and resuspended into 100 µl TE.

2.8.2 Plasmid preparation by the alkaline lysis method.

The method used is essentially as described by Maniatis et al. (1982). 10 ml of LB medium (+ suitable antibiotic) was innoculated with a single colony and grown overnight at 37°C. 2.5 ml of this culture was added to a two-litre flask containing 500 ml LB medium (+ antibiotic) and grown overnight at 37°C. The culture was separated into two 250 ml centrifuge tubes (GSA) and centrifuged at 6000 rpm for 10 minutes (4°C). Each pellet was resuspended in 25 ml TGE (25 mM Tris-HCl pH 8.0, 50 mM glucose, 10 mM Na_2EDTA), 5 ml lysozyme (10 mg/ml in TGE) was added and the bottles were stood at room temperature for 10 minutes. 60 ml of 0.22 M NaOH/10% SDS (9 : 1) was added, mixed carefully and the bottles were stood on ice for 5 minues. 30 mls ice-cold 5 M potassium acetate (pH 5.0) was mixed in and the bottles stood for 15 minutes on ice. Supernatants from one 10 minute spin at 6000 rpm were filtered through gauze to remove debris and 0.6 volumes of ispropanol used to precipitate DNA. The precipitated DNA was pelleted by centrifugation for 10 minutes at 8000 rpm (4°C) and resuspended in 8 ml TE buffer. pH was brought to neutrality with a few drops of 3 M Tris. Phenol extraction and ethanol precipitation (section 2.6.2) gave cleaned up DNA in 2 ml TE. 4.9 g CsCl and 250 µl of 10 mg/ml ethidium bromide were added and the volume was brought to 5.6 ml. The supernatants were sealed in 13 x 51 mm "Quickseal" Beckmann tubes and centrifuged overnight in a VTi65 rotor at 50,000 rpm (20°C) to band the

plasmid DNA (two bands are seen, the upper band is bacterial genomic DNA, the lower one (banding to the middle of the tube, is the plasmid band). The banded plasmid was visualised under long wavelength ultraviolet (Ultraviolet Products Inc, Cambridge, England) and the band was removed with syringe needles. N-butanol ($CsCl/H_2O$-saturated) extraction was performed to remove ethidium bromide. The DNA was ethanol precipitated and resuspended in 500 µl TE.

2.8.3 Single-stranded M13 DNA preparation

1.5 ml LB medium was innoculated with 15 µl of saturated, overnight growth JM101 and single plaques were picked into each tube using a toothpick. Growth by vigorous shaking at 37°C for 4.5-5.5 hours was followed by transfer to eppendorf tubes and centrifugation for 5 minutes at maximum speed. Supernatants were removed to new tubes and 150 µl PEG (2.5 M NaCl, 20% PEG 6000) was added. This was mixed, then placed at 4°C for 10 minutes, then spun for 5 minutes in a microfuge at maximum speed. The supernatant was removed and the pellet resuspended in 100 µl TE. Phenol extraction and ethanol precipitations were performed as usual and the DNA was resuspended in a final volume of 15-20 µl. Single stranded DNA templates are produced, recombinants can be identified due to slower migration relative to the M13 vector, by running on agarose gels. The single-stranded template is used for M13 DNA sequencing (see section 2.11).

2.8.4 Preparation of genomic DNA

For preparing DNA from blood lymphocytes, 5 ml of blood was incubated with 10 ml lysis solution (155 mM NH_4Cl, 10 mM $KHCO_3$, 0.1 mM Na_2EDTA) on ice for 15 minutes and was then spun at 3000 rpm for 10 minutes at 4°C. The pellet of

white cells was resuspended in 10 ml SE (75 mM NaCl, 2 mM Na_2EDTA), vortexing to resuspend cells. The cells were again centrifuged for 10 minutes and resuspended in 5 ml SE buffer. 25 µl proteinase K (20 mg/ml in SE) and 1 ml 10% SDS were added, mixed in and incubated at 37°C, overnight. Phenol extraction was performed, followed by isopropanol precipitation of DNA (200 µl 3 M Na-Acatate pH 5.0 and 6 ml isopropanol were added and mixed by repeated inversion). The DNA precipitates as a string-like mass and can be hooked out using flamed pasteur pipettes. These were allowed to air-dry and eased off the hook into 500 µl TE, heated at 65°C for 10 minutes and placed at 4°C overnight, to resuspend.

Liver and mammary DNA was prepared from frozen tissue. 0.25-0.5 g of frozen tissue was added to 4 ml 10 mM Na_2EDTA, 50 mM Tris-HCl pH 8.0, and homogenised for a few seconds. 224 µl 4 M NaCl, 60 µl proteinase K (as above) were added and mixed gently. Then 1.2 ml 10% SDS was added and this mixture was incubated at 37°C, overnight, with gentle shaking. Two phenol extraction and two phenol/chloroform extractions were performed. DNA was precipitated by isopropanol addition (as above).

Both methods yielded 1000-3000 µg DNA per g tissue.

2.8.5 RNA preparation by the guanidinium thiocyanate/CsCl gradient method

RNA isolation was performed according to the protocol of Glisin et al. (1974) and Chirgwin et al. (1979). About 0.5 g of tissue was homogenised in 4.5 ml 4 M guanidinium thiocyanate (prepared as follows: 25 g guanidinium thiocyanate, 0.25 g sodium N-laurylsarcosine, 1.25 ml 1 M sodium citrate, 0.165 ml sigma

antifoam A (0.1%), made up to 50 ml and pH 7.0 with 1 N NaOH. Add 1.25 ml 0.2 M DTT). 4 ml of the homogenate was layered onto 2 ml of 5.7 M CsCl/25 mM Na-Acetate and centrifuged in 13 x 50 mm Beckman Ultra-Clear tubes at 36,000 rpm (Beckman SW50.1 rotor) for 12 hours (20°C). RNA comes through the CsCl cushion and pellets at the bottom of the tube whilst DNA and protein remains above the cushion.

The supernatant was carefully removed after centrifugation, the pellet was resuspended in 1 ml 7.5 M guanidinium HCl (prepared as follows: 35.82 g guanidinium hydrochloride, 1.25 ml 1 M sodium citrate, made up to 50 ml and pH 7.0 with 1 N NaOH. Add 1.25 ml 0.2 M DTT) and precipitated with 0.025 ml 1 N acetic acid and 0.5 ml ethanol, at 20°C, overnight. The RNA was pelleted by spinning at 10,000 rpm for 20 minutes (4°C) and resuspended in distilled water. Two ethanol precipitations using Na-acetate and 2-3 volumes of ethanol were performed before final resuspension in 500-1000 µl.

Yields of RNA varied from 400 µg to 4000 µg per g tissue.

Amounts of DNA and RNA obtained by the above methods were determined by reading absorbance at 260 and 280 nm and by running small aliquots on agarose gels, together with concentration markers.

2.9 DNA/RNA TRANSFER TO MEMBRANES

2.9.1 Southern transfer

Amersham Hybond nylon membranes were used for all Northern and

Southern transfers. DNA gels were immersed and gently shaken in 0.25 M HCl for 2 x 15 minutes to depurinate, followed by 2 x 20 minutes in 1.5 M NaCl, 0.5 M NaOH (denaturation step). The neutralisation step used 2 x 25 minutes in 0.5 M Tris-HCl, 1.5 M NaCl, pH 7.4. The gel was then rinsed in 2 x SSC (20 x SSC was made with 175.3 g NaCl, 88.2 g tri-sodium citrate to a final volume of 1 litre, after pHing to 7.0 with HCl or NaOH, as appropriate) and placed on a transfer apparatus, as described by Maniatis et al. (1982). 20 x SSC was used in the reservoir (and wick). The Hybond nylon membrane does not require wetting before it is placed on the gel. Wet paper towels were replaced by dry ones 30 and 60 minutes after the transfer had been set up and then left overnight.

After removal from the transfer apparatus, the DNA/RNA filter was rinsed in 50 mM NaPi, pH 7.2, air-dried (making sure the filter did not completely dry out) and was wrapped in "Saran Wrap". The DNA/RNA was irreversibly fixed to the membrane by UV irradiation (for 45 seconds on our trans-illuminator).

2.9.2 Northern transfer

Depurination or denaturation is not required for RNA transfer. After electrophoresis, therefore, the gel was rinsed in 1 x SSC and set up for transfer, as described in Maniatis et al. (1982). 10 x SSC was used in the reservoir. 1 x SSC was used to wet the 3 MM paper which was placed on the membrane. The remainder of the procedure was the same as that used for DNA transfer (above). RNA gels which were transferred were not stained with ethidium bromide.

2.10 HYBRIDISATION

2.10.1 Oligo-labelling

The method of Feinberg and Vogelstein (1983, 1984) was used to make DNA probes for hybridisation. The DNA must be linearised by restriction digestion for good labelling. Up to 100 ng of restriction digested DNA was diluted to 32.5 µl with H_2O, placed in a boiling water bath for 3 minutes to denature the DNA and placed at 37°C for 10-60 minutes. 10 µl of OLB buffer (described by Feinberg and Vogelstein, 1984), 2 µl bovine serum albumin (10 mg/ml), 5 µl of high specific activity (~3000 Ci/mmol) [α ^{32}P] dCTP and 1 unit of *E. coli* DNA polymerase I Klenow fragment, were added. Labellings were incubated at room temperature for 5 hours or overnight. Cerenkov counting of trichloroacetic acid (TCA) precipitated labelled DNA was done to check ^{32}P incorporation. 1 µl of the oligo-labelled preparation was removed into another tube. 2 µl bovine serum albumin (10 mg/ml) and 200 µl 5% TCA were added, vortexing to mix. This was passed over a Whatman GF/C 2.5 cm filter and washed with more 5% TCA, to remove all unincorporated label. The filter was Cerenkov counted for ^{32}P incorporation.

2.10.2 Hybridisation

DNA and RNA filters were hybridised using a protocol adapted from Church and Gilbert (1984). Hybridisation filters were rinsed in 50 mM NaPi (pH 7.2) and placed in hybridisation bags cut just larger than the size of the filter(s). 30-40 ml prehybridisation buffer (0.5 M NaPi pH7.2, 7% SDS, 1 mM Na_2EDTA) was added, the bag was sealed after removal of air bubbles and placed at 65°C, on a shaking platform, for 5-20 minutes. The probe was denatured by addition of NaOH to 0.4 M

and was added to prehybridisation buffer (now hybridisation buffer). The prehybridisation bag was opened on one side, buffer was removed, hybridisation buffer was added, the bag was re-sealed and placed on a shaking platform at 65°C, overnight. Up to 1.5 x 10^6 Cerenkov counts were added per ml of buffer.

40 mM NaPi pH 7.2, 1% SDS was used to wash filters after hybridisation. Filter(s) were washed with three 5 minute washes and one 20 minute wash, all at 65°C. The filters were air-dried, wrapped in Saran wrap and autoradiographed.

2.11 DNA SEQUENCING

Sequencing of M13 recombinant clones was carried out according to the methods of Sanger et al. (1978) and Biggin et al. (1983) using [α ^{32}P] dATP or [α ^{35}S] dATP. M13 recombinant clones were prepared as described in section 2.6.2. 5 µl of M13 DNA, 2 µl primer (Pharmacia M13 17-mer at 0.03 A_{260} units), 2 µl TM buffer (100 mM Tris-HCl pH 8.0, 100 mM $MgCl_2$), 1 µl H_2O were mixed in an eppendorf tube, placed in a 70°C water bath for 3 minutes and allowed to cool to room temperature over 90 minutes.

The appropriate "NTP mixes", sufficient for 50 reactions, were prepared as below:

NTP mixes for sequencing

	G	A	T	C
0.5 mM dGTP	2.5	25	25	25
0.5 mM dTTP	25	25	2.5	25

0.5 mM dCTP	25	25	25	2.5
10 mM ddGTP	3.5			
10 mM ddATP		0.25		
10 mM ddTTP			4	
10 mM ddCTP				1.5
TE buffer	50	50	50	50

Sequencing reactions were carried out in four tubes labelled G, A, T and C, into which the various components were added on the sides of the tubes and were mixed by quick spins in a microfuge. All amounts above and below are in µl.

	G	A	T	C
"G" mix	2			
"A" mix		2		
"T" mix			2	
"C" mix				2

2 µl of the annealed clone was now added to each tube and 2 µl of Klenow mix (1 µl [α ^{35}S] dATP (10µCi/µl; >1000Ci/mmol), 0.8 µl Klenow (4.5 units/ml), 7.2 µl TE) was added. This was mixed by spinning in a microfuge and placed at 37°C for 20 minutes. After 20 minutes the reaction was "chased" by adding 2 µl of the chase mix (0.25 mM solution of each dNTP made up in TE). This was mixed by spinning and placed at 37°C for 35 minutes. 4 µl of dye (100 ml deionised formamide, 0.1 g xylene cyanol FF, 0.1 g bromophenol blue, 2 ml Na_2EDTA) was added to stop the reaction. The reactions can now be kept at -20°C for several weeks. The reactions were placed in a boiling water bath for 3 minutes prior to loading. 2-3 µl of each reaction was loaded using shark's tooth combs.

2.11.1 Acrylamide gels

50 ml volume 8% acrylamide, 7 M urea gels were prepared (21 g urea, 10 ml 40% acrylamide (19:1 acrylamide: N, N', methylene-bisacrylamide), 2.5 ml 20 x TBE, pH 8.8 (300 g Tris base, 50 g boric acid, 20 g Na_2EDTA, made up to 1 litre) and 21.5 ml H_2O. This was stirred until the urea was in solution and degassed. 300 µl fresh 10% (w/v) ammonium persulphate and 25 µl TEMED were added and the gel was poured between 20 x 36 cm plates, a shark's tooth comb was inserted and the gel was allowed to set for 1 hour.

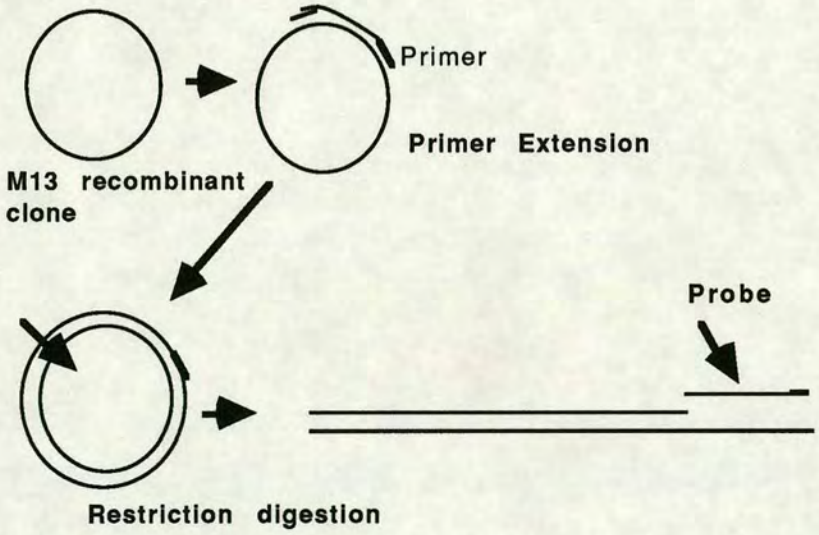
The gel was set up for vertical electrophoresis using 1 x TBE as running buffer. The gel was pre-run, at 30 mA, for 15-20 minutes. The comb was taken out, the comb space was flushed out with a syringe to remove urea which leaches out of the gel. The comb was placed and samples loaded. The gels were run at 30 mA, for 2-8 hours, fixed in 10% acetic acid, dried on a Biorad gel dryer and autoradiographed overnight at room temperature, without intensifying screens.

2.12 S1 MAPPING

S1 mapping was carried out using the protocol described by Davis et al. (1986). This involves the use of M13 single-stranded DNA to generate the probe by primer extension, digestion with a restriction enzyme which will cut once within, or just outside the probe DNA. The probe preparation scheme is shown diagrammatically below.

The annealing reaction was carried out as described above in section 2.11. 5 µl of M13 DNA (prepared as described in 2.8.3) should be sufficient for about 10

reactions. Once annealing was complete 4 µl C^{o} (2 mM dATP, dGTP, dTTP), 1 µl 0.1 M DTT, 4 µl [α ^{32}P] dCTP (10 µCi/µl; 3000 Ci/mmol) and 1 µl Klenow fragment (4.5 unit/µl) were added, mixed by spinning and incubated at room temperature for 10 minutes. 4 µl Chase (0.25 mM dNTPs) was added and incubated at room temperature for a further 10 minutes.



H_2O was added to a final volume of 100 µl. Phenol/chloroform extractions and ethanol precipitation were performed and the DNA was resuspended in 20 µl 1 x appropriate restriction buffer (in the case described in chapter 4 an *EcoRI* digestion was performed). A 1 hour digestion was performed, followed by phenol/chloroform extractions and ethanol precipitation.

These steps lead to the final situation depicted in the figure above. Separation on a denaturing gel and isolation of the small fragment (M13-derived S1 probes cannot be much larger than 2 kb in size since no more DNA can be ligated into M13, due to limits of phage packaging). The S1 probe, described in chapter 4, was

about 170 nucleotides long, so isolation from acrylamide gels was performed.

An 8% urea-acrylamide gel was prepared (as described above). The restricted DNA was resuspended in 5 µl TE buffer. 5 µl formamide dye (described in section 2.11) was added and the sample was placed in a boiling water bath for 10 minutes. The whole sample was loaded and the gel was run at 30 mA for 2 hours. The gel was wrapped in Saran Wrap and a X-ray film was placed on top and marked. A 2 minute exposure was usually sufficient for positions of the three expected fragments to be determined. The region containing the probe fragment was cut out, 600 µl of 300 mM NaCl, 1 mM Na_2EDTA, 10 mM Tris-HCl pH 8.0, was added and the fragment was allowed to diffuse out at room temperature, overnight. The supernatant was removed after spinning for 10 minutes and phenol/chloroform extractions and ethanol precipitation performed.

The above method yields a single-stranded DNA probe, so DNA-DNA hybrid formation is not a problem during S1 digestion.

The pellet was resuspended in 300 µl hybridisation solution (3000 µl deionised formamide, 400 µl 10 x hybridisation buffer (200 mM Tris-HCl pH 7.4, 4 M NaCl, 10 mM Na_2EDTA, 0.1 M DTT), 40 µl 10% SDS, 60 µl H_2O). 1 µl was Cerenkov counted and the volume adjusted to 25,000-100,000 cpm/35 µl.

RNA to be analysed was ethanol precipiated and pelleted. 10 µg was used, per reaction. The pellet was resuspended in 35 µl of the above hybridisation solution and incubated at 65°C, overnight. The hybridisation temperature was estimated from G+C content and from testing different hybridisation temperatures.

1 µl salmon sperm DNA (2 mg/ml), 1.5 µl S1 nuclease (33 units/µl), 280 µl H_2O, 80 µl 5 x S1 buffer (3 ml 5 M NaCl, 1.66 ml 100 mM $ZnSO_4$, 1 ml 3 M Na-acetate pH 4.5, 4.3 ml H_2O) were mixed and 350 µl was added to each

hybridisation. S1 digestion was performed at 37°C for 1 hour. Phenol/chloroform and ethanol precipitation (no salt added) were performed and the pellet resuspended in 3 µl TE.

The S1 digest samples were run on 8% urea-acrylamide gels for 2 hours and autoradiographed, with screens, at -70°C.

A number of control digests were performed. These were S1 digestion of the probe after hybridisation with *Saccharomyces cerevisiae* tRNA and incubation of the probe with the S1 buffer, but without S1 being added.

2.13 ISOELECTRIC FOCUSING

2.13.1 Preparation of Milk samples

Milk samples were acid-treated to remove caseins and enrich for whey proteins (see Simons et al., 1987). 100 µl of each milk sample was diluted to 500 µl. This was given a 5 second spin to bring fat to the top. The milk was taken off, leaving the fat. 1 N HCl was added, pH checked and more HCl was added, until pH was down to 4.5-4.6, at which point the caseins precipitate out. This was spun for 5 minutes in a microfuge at maximum speed and the supernatant was removed to a new tube. 100 µl of this was dialysed against H_2O, at 4°C, to bring pH back to neutrality.

2.13.2 Isoelectric Focusing Polyacrylamide Gels

Pharmacia-LKB poured flat bed, ampholine polyacrylamide gels for pH 3.5 to 9.5, were used for isoelectric focusing. The procedure used was as described in the LKB leaflet. The gel was placed on a cooling plate after a thin layer of an insulating

had helped
fluid (kerosene) been spread over the plate. Circulating water to maintain
were
a low and constant temperature across the entire plate. Two electrode strips
were
soaked in the anode solution (1 M H_3PO_4) or the cathode solution (1 M NaOH) and
were
placed across the gel, close to the edges, along the length of the gel. These strips
in contact with the electrodes when running.15 µls of the dialysed whey preparations were applied to application strips and the strips were placed along the gel parallel, and close to, the cathode strip. A pH gradient forms across the gel, from the anode to the cathode. The proteins leave the strips and migrate to their pIs, where they remain. The gels were run at 1500 V, 50 mA, 30 W. When first connected the voltage was approximately 700 V. Throughout the run voltage climbs, reaching a plateau at 1500 V, whilst current falls to a low of about 10 mA. Once maximum
was
voltage has been reached (in approximately 2 hours) the gel is run for a further 10 minutes, then removed. The gels were fixed in trichloroacetic acid/sulphosalicycylic acid (57.5 g TCA, 17.25 g sulphosalicylic acid in 500 ml distilled water) for 30-60 minutes. This allows the ampholines to diffuse out and precipitates the proteins. The gels were then placed in destaining solution (500 ml ethanol, 160 ml acetic acid diluted to 2 litres with distilled water) for five minutes, then moved to 400 ml staining solution (0.46 g Coomassie Blue R 250 dissolved in 400 ml destaining solution) at 60°C. The gels were destained with several changes of destaining solution and photographed (see figures 5.2a and 5.5).

2.14 COMPUTING

All
Various computing facilities were used in the course of this work. the

DNA and protein sequence manipulation and analysis was done using the University of Wisconsin package of programs (Devereux et al., 1984), with the exception of the tree generation programs which were devised by J. Felsenstein (see chapter 6). The EMBL, GENBANK and NBRF databases were extensively used for searches. All these programs were run on the Edinburgh University and AFRC VAX computers.

This thesis was written on an Apple Macintosh SE microcomputer.

Chapter 3 EXPRESSION OF OVINE MILK PROTEIN GENES DURING PREGNANCY AND EARLY LACTATION

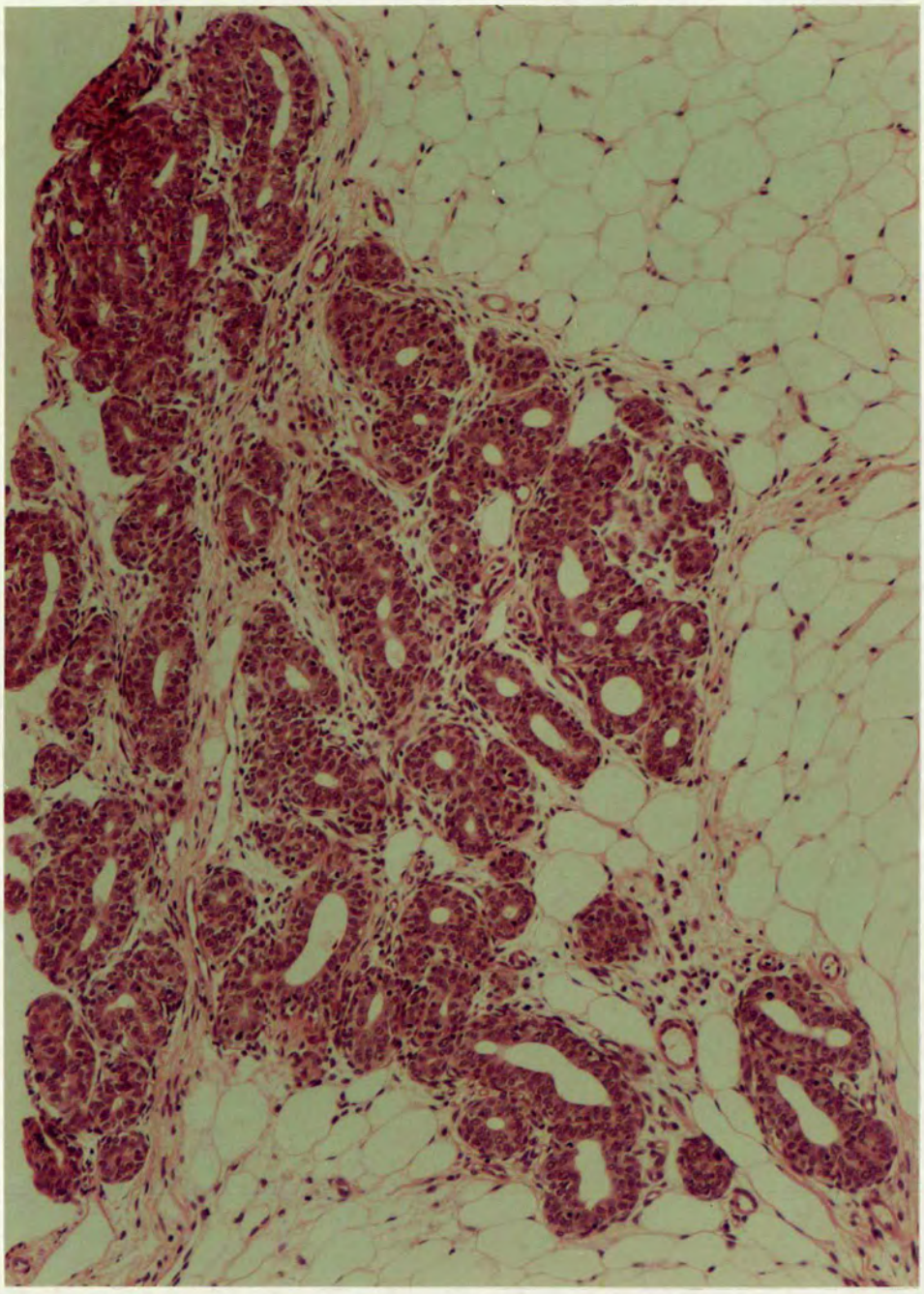
Mammary gland development is hormonally regulated, largely by the ovarian steroid hormones, progesterone and oestrogen and the pituitary hormone, prolactin. In most mammals much of the mammary gland development occurs during pregnancy, although in some mammals a great deal of mammary gland development occurs after parturition (for example, in the rat and rabbit - 50% and 30% of total mammary development, respectively, occurs soon after parturition). In sheep, 78% of mammary growth occurs during the first pregnancy and only 2% during lactation (Anderson, 1975; Forsyth, 1983). Chapter 1 describes mammary gland development and also describes work which has shown that milk protein gene expression is hormonally regulated during pregnancy and that milk protein gene expression in non-coordinately regulated. Prolactin and glucocorticoids are the main lactogenic hormones but progesterone is important for correct regulation of milk protein gene expression during pregnancy. It inhibits milk protein gene expression and falls in levels of progesterone towards the end of pregnancy appear to initiate milk protein production (see chapter 1).

Increases in milk protein gene expression occur at a specific stage in many mammals. This stage can be identified histologically by increases in cell number, appearance of lactose and rapid increases in DNA and RNA levels (see Denamur, 1974). In the rabbit, significant increases in wet weight and nucleic acid content occur between days 16-24 of pregnancy (in rabbits, the gestational period is about 30 days). Lactose synthetase activity and secretion of milk first occurs during this period (see Shuster et al., 1976). Shuster et al. (1976) showed that casein mRNA levels

increase greatly at this period. In the mouse, milk protein synthesis becomes detectable at days 12-15 of pregnancy (see Denamur, 1974) and β-casein mRNA levels increase greatly at day 12 of gestation (S. Harris, unpublished results) whilst WAP mRNA levels rise from day 16 (S. Harris, unpublished results; Pittius et al., 1988); the gestational period is 18-21 days in the mouse. In the rat, mammary RNA levels increase throughout pregnancy (the gestational period in the rat is about 21 days), following increases in DNA content (see Denamur, 1974). Analysis of mammary casein and WAP levels shows that protein and mRNA levels increase throughout pregnancy, rapid increases occurring after parturition. Nakhasi and Qasba (1979), however, presented results showing major increases in rat α-lactalbumin mRNA levels from day 11 of pregnancy. Thus, in many mammals (see sheep below) major increases in levels of milk protein mRNAs occur at the stage of pregnancy, just past mid-pregnancy, when cell number increases rapidly and milk secretion first becomes detectable in histological examinations.

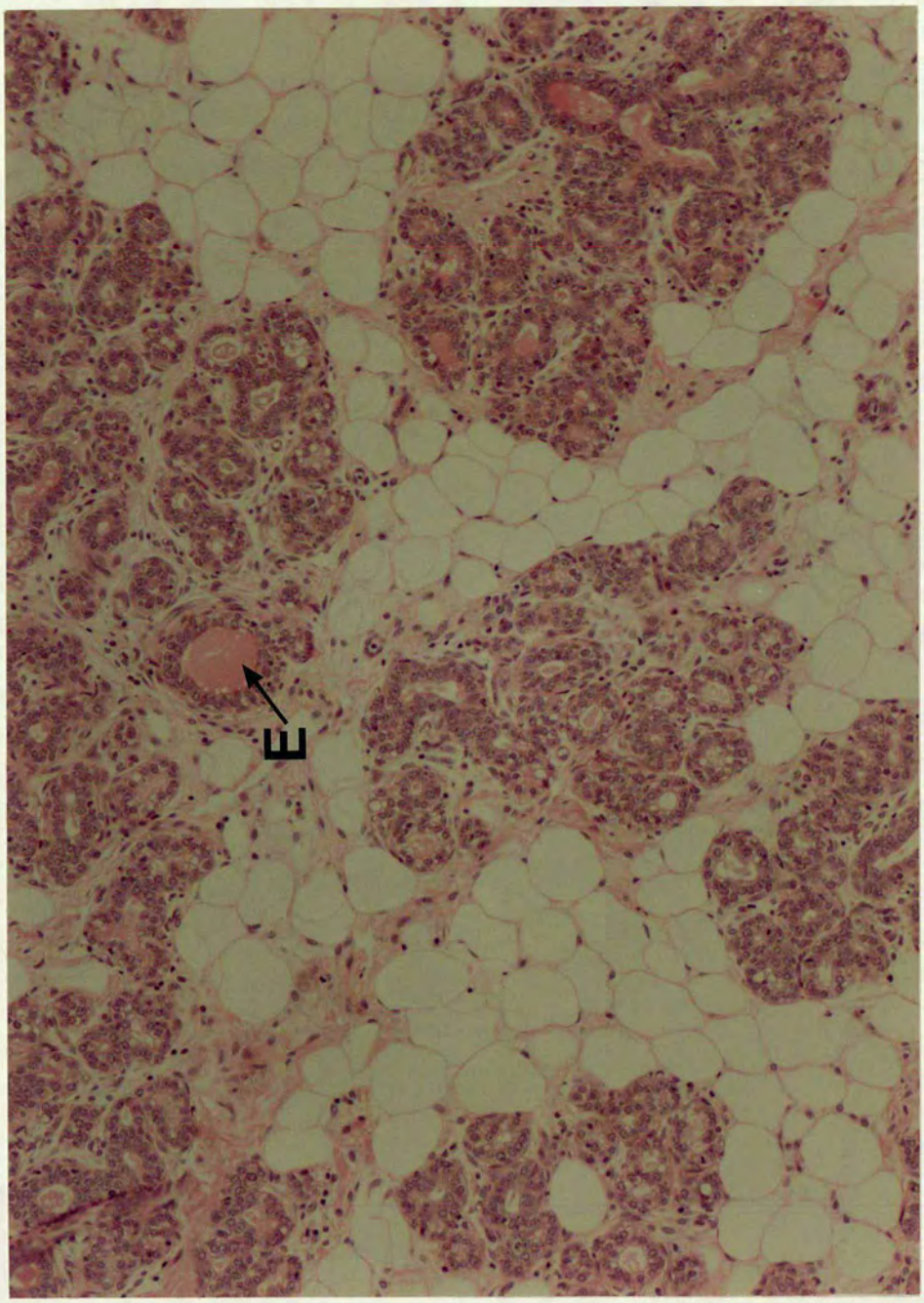
In sheep, Denamur and colleagues (see Denamur, 1974) have found that mammary growth is slow between days 0-90 of pregnancy (the gestational period is about 148 days in sheep). During this period the DNA content increases slowly. Increase in RNA content per mammary gland is also slow during this period. Histological data and lactose content analysis shows that secretion begins at day 90 of pregnancy (see Denamur, 1974). At this stage (days 90-145 of pregnancy) DNA content increases at a greater rate and RNA content also increases rapidly. RNA/DNA ratios increase after day 90 and peak at 21 days of lactation (see Denamur, 1974; and references therein). During lactation 60-80% of poly(A)$^+$ RNA encodes milk protein genes in the ewe, rabbit, rat, guinea pig, man and cow (see Mercier and Gaye, 1983).

a



b

C



d

e

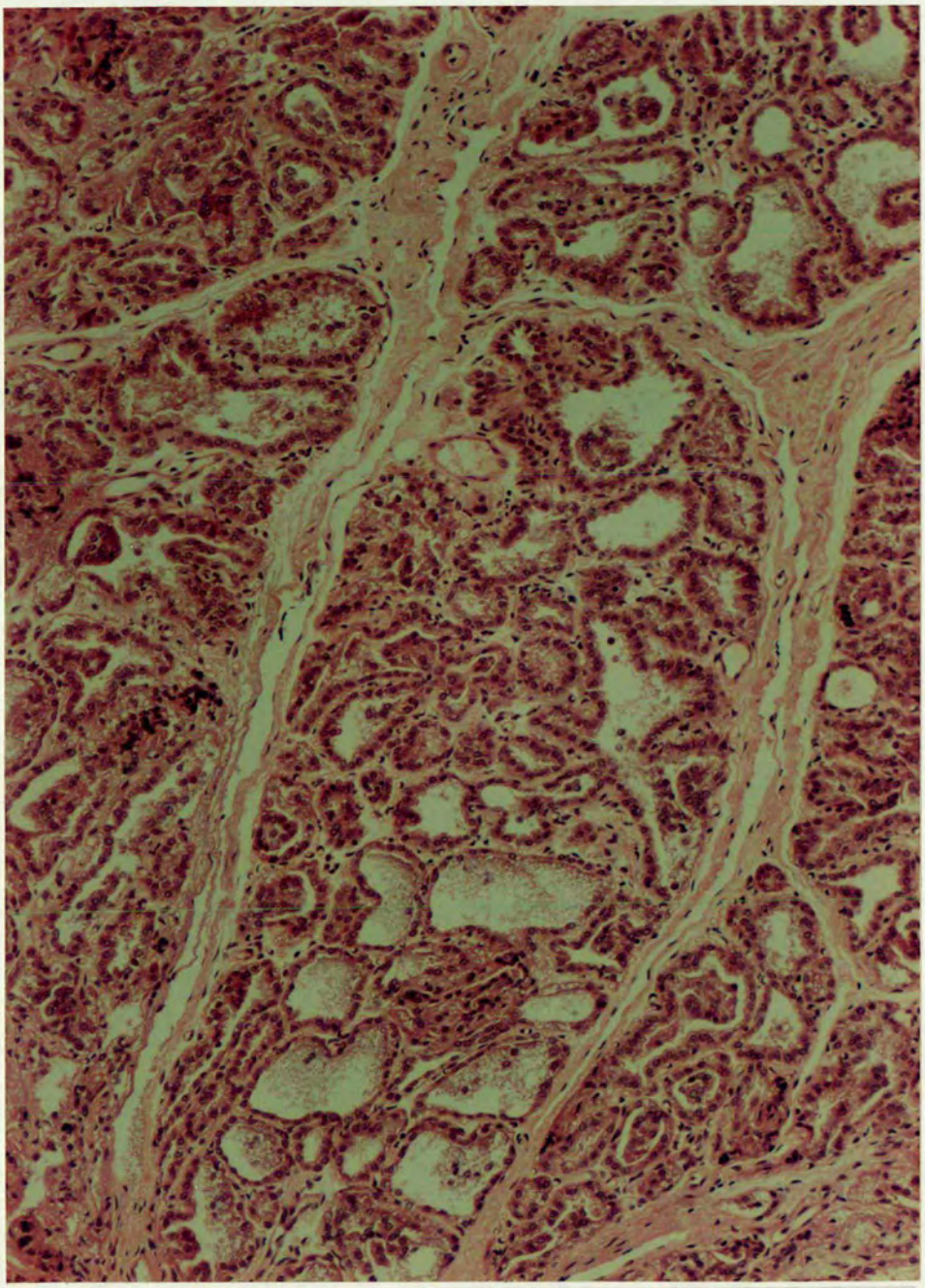


Figure 3.1. Histological examination of sheep mammary gland development during pregnancy. Mammary gland tissue was fixed in Tellyesniczky's fluid, 7 μm section were cut and stained with haemotoxylin and eosin (H&E) (see Dils and Forsyth, 1981). (a) virgin, (b) 74-days pregnant, (c) 90-days pregnant, (d) 120-days pregnant, 145-days pregnant and, (f) 21-days lactating gimmer mammary gland cross-sections are shown. The presence of exudate (E) is evidenced by the strong staining region in the lumen from day 90 of gestation. Fat globlules appear in the lumen from day 120. See Anderson (1975) and Fleming et al. (1986) for details of mammary gland structure. Magnification, x 184. These figures were kindly provided by Maggie McClenaghan.

In order to examine milk protein gene expression and in particular, to follow BLG gene expression, ewes were sacrificed at various stages of pregnancy. Virgin, first pregnancy (called gimmer) ewes and fourth pregnancy (called draft) ewes were used in the time-course. Non-pregnant, 5-day, 20-day, 70-day, 90-day, 110-day and 145-day pregnant, post-partum, 1-day, 5-day, and 20-day lactating, draft ewes were sacrificed. Mammary and liver tissue was flash-frozen in liquid nitrogen and some samples were homogenised in guanidinium thiocyanate for preparation of total RNA (see Materials and Methods). A more extensive time-course was designed for first pregnancy ewes, particularly concentrating around days 75-120, as major changes have been shown to occur within this period (see Denamur, 1974). In addition to preparation of RNA, mammary gland samples were fixed in Tellyesniczky's fluid (see Dils and Forsyth, 1981) for histological analysis (M. McClenaghan). Some of these sections are shown in figure 3.1.

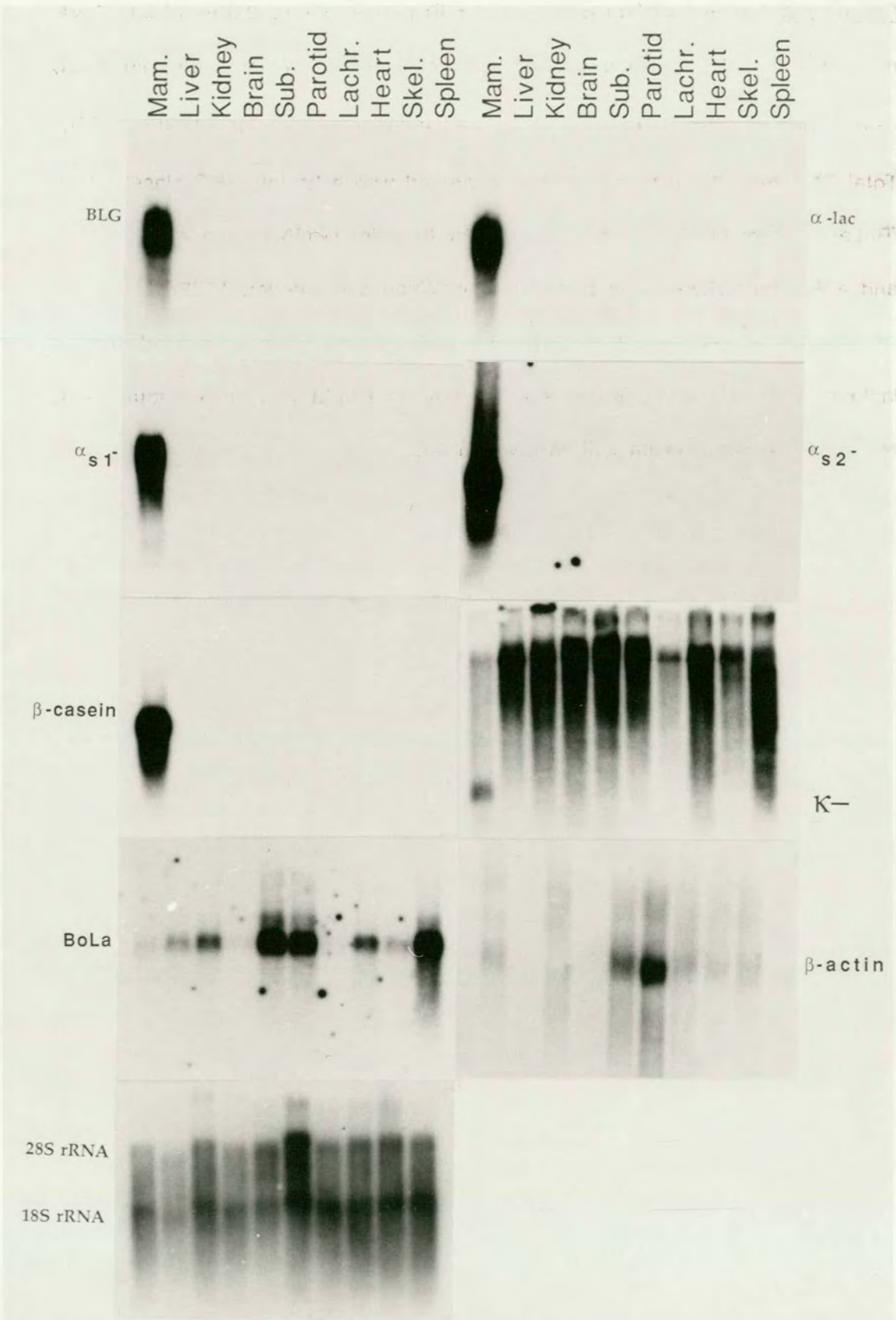
In agreement with the results described by Denamur (1974), histological sections show evidence of secretory activity in the alveolar lumen at day 90 of gestation (figure 3.1c). Prior to this stage little change in mammary gland structure is seen (see figure 3.1a, b). A greater cell number and more secretory activity is apparent at day 100 (no cell counts have been made for this time-course). Between days 90-120, milk secretion increases, enlarging the lumen. First evidence of "globules" is seen at day 120 (figure 3.1d) and by day 145 (figure 3.1e) much of the lumen volume is taken up by these globules. These are probably fat globules, their sizes are in the range described by Walstra and Jenness (1984). The 145-day pregnant mammary gland has an appearance similar to that of the post-partum mammary gland. The 21-day lactating mammary gland (figure 3.1f) is similar to the 145-day pregnant mammary gland, showing a similar distended lumen structure, and a similar "secretion" in the lumen. From the appearance of secretory activity in the

mammary gland, it appears that day 90 and day 145 are key stages in sheep mammary gland development. Furthermore, the results of this histological analysis are consistent with changes in mammary gland morphology described previously (see above). These sections were processed and photographs were taken by M. McClenaghan, whose permission to use them to illustrate changes in the ovine time-course, is gratefully acknowledged.

3.1 TISSUE-SPECIFIC EXPRESSION OF SHEEP MILK PROTEIN GENES

In order to show that the milk protein genes are tissue-specific, a mid-lactational ewe was sacrificed and total RNA was prepared from ten tissues, including the mammary gland, salivary gland (parotid and submaxillary glands), lachrymal gland, liver and kidney. 2.0 μg of each RNA was run on 1% formaldehyde-MOPS agarose gels (see Materials and Methods) and transferred to nylon membranes. cDNAs for ovine BLG, α-lactalbumin, α_{s1}-, α_{s2}-, β-, and κ-caseins were used to probe the Northern filters. These cDNAs were obtained from J-C. Mercier (Mercier et al., 1985; Gaye et al., 1986; 1987). The ovine RNAs were also probed with three non-tissue-specific gene probes, to show that the RNAs were not degraded and contain RNA species which should be present in these tissues. A human β-actin cDNA containing 3' untranslated sequences (Ponte et al., 1983), a cDNA encoding a bovine MHC class I (BoLa) gene (provided by P. Brown; Brown et al., 1989) and a

Figure 3.2. Tissue-specific expression of milk protein genes. 2.0 μg of total RNA from the ten tissues was probed with cDNAs for the milk protein BLG, α-lactalbumin (α-lac), α_{s1}-, α_{s2}-, β- and κ-caseins (Mercier et al., 1985). Total RNA from these tissues was also probed with a bovine MHC classI cDNA (BoLaI - Brown et al., 1989), a β-cytoactin (β-actin) cDNA (Ponte et al., 1983) and a Xenopus rDNA gene probe (Sollner-Webb and Reeder, 1979). Mammary (mam.), liver, kidney, brain, submaxillary gland (sub.), parotid gland, lachrymal gland (lachr.), heart, skeletal muscle (skel.) and spleen total RNA, prepared from a mid-lactational ewe, was used.


Mam.
Liver
Kidney
Brain
Sub.
Parotid
Lachr.
Heart
Skel.
Spleen
Mam.
Liver
Kidney
Brain
Sub.
Parotid
Lachr.
Heart
Skel.
Spleen
BLG
α-lac
αs1-
αs2-
β-casein
κ-
BoLa
β-actin
28S rRNA
18S rRNA

Xenopus rDNA gene (Sollner-Webb and Reeder, 1979) were used as probes.

Figure 3.2 shows that ovine BLG, α-lactalbumin, α_{s1}-, α_{s2}- and β-casein genes are only expressed in the mammary gland. Longer exposures did not show the appearance of low abundance mRNAs in any other tissue. Probing with the κ-casein cDNA showed the presence of a mammary-specific RNA species of the expected size for ovine κ-casein. In addition, however, a "smear", absent from mammary RNA, was seen in all other tissues. The correct-sized band is absent from all tissues apart from the mammary gland. It is possible that the cloned cDNA contains a large poly(A) tail which hybridises with all poly(A)$^+$ RNA. Alternatively, the κ-casein cDNA may contain a repeat element which forms part of the transcription unit of some genes. Certainly, the lower abundance of this signal in the lachrymal gland does not reflect the levels of RNA loaded (compare with the BoLa, β-actin mRNAs and rRNA levels observed in this sample), suggesting that the smear is not due to hybridisation against poly(A). Furthermore, WORDSEARCH (Devereux et al., 1984) comparison of the 3' untranslated sequences of the ovine κ-casein cDNA sequence against the GENBANK database suggests sequence similarity with a human interspersed, repetitive sequence (Pan et al., 1981) (the fifth best match; the first three were cow, rat and mouse κ-casein cDNA sequences, respectively). The absence of a smear from mammary RNA may simply reflect the fact that RNA polymerase II-transcribed RNAs consist largely of a few, abundant species (the milk protein genes), so that mRNAs containing this repeat (or a poly(A) tail) are far fewer in the mammary gland, relative to total RNA, resulting in its apparent absence from the mammary gland RNA.

Probing for β-actin and BoLa mRNAs and rRNA shows that RNA from all ten tissues are intact and not degraded. Amounts of RNA used were determined from optical density readings and should be similar loadings. Differences in signal, therefore, probably reflect differential expression of these genes in the ten tissues. Thus,

Xenopus rDNA gene (Sollner-Webb and Reeder, 1979) were used as probes.

Figure 3.2 shows that ovine BLG, α-lactalbumin, α_{s1}-, α_{s2}- and β-casein genes are only expressed in the mammary gland. Longer exposures did not show the appearance of low abundance mRNAs in any other tissue. Probing with the κ-casein cDNA showed the presence of a mammary-specific RNA species of the expected size for ovine κ-casein. In addition, however, a "smear", absent from mammary RNA, was seen in all other tissues. The correct-sized band is absent from all tissues apart from the mammary gland. It is possible that the cloned cDNA contains a large poly(A) tail which hybridises with all poly(A)$^+$ RNA. Alternatively, the κ-casein cDNA may contain a repeat element which forms part of the transcription unit of some genes. Certainly, the lower abundance of this signal in the lachrymal gland does not reflect the levels of RNA loaded (compare with the BoLa, β-actin mRNAs and rRNA levels observed in this sample), suggesting that the smear is not due to hybridisation against poly(A). Furthermore, WORDSEARCH (Devereux et al., 1984) comparison of the 3' untranslated sequences of the ovine κ-casein cDNA sequence against the GENBANK database suggests sequence similarity with a human interspersed, repetitive sequence (Pan et al., 1981) (the fifth best match; the first three were cow, rat and mouse κ-casein cDNA sequences, respectively). The absence of a smear from mammary RNA may simply reflect the fact that RNA polymerase II-transcribed RNAs consist largely of a few, abundant species (the milk protein genes), so that mRNAs containing this repeat (or a poly(A) tail) are far fewer in the mammary gland, relative to total RNA, resulting in its apparent absence from the mammary gland RNA.

Probing for β-actin and BoLa mRNAs and rRNA shows that RNA from all ten tissues are intact and not degraded. Amounts of RNA used were determined from optical density readings and should be similar loadings. Differences in signal, therefore, probably reflect differential expression of these genes in the ten tissues. Thus,

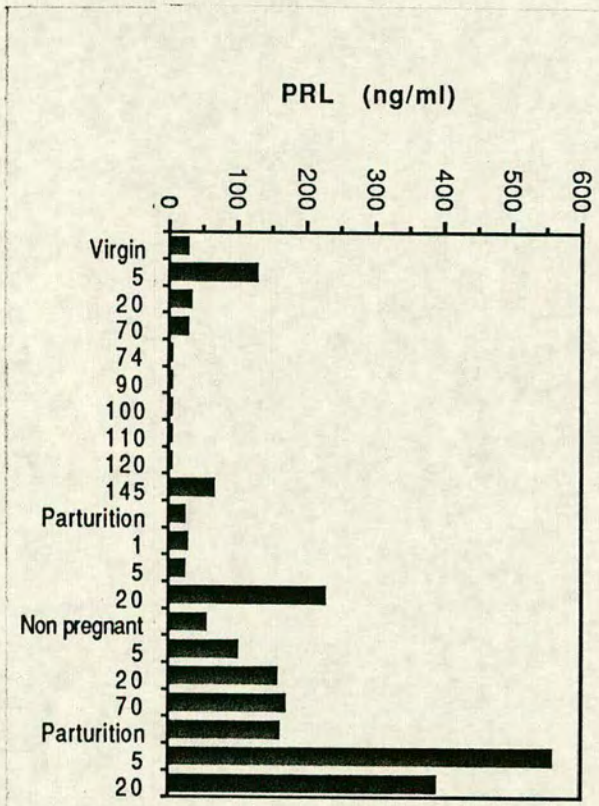
expression of the MHC Class I gene is highest in the spleen where lymphocytes are abundant. MHC Class I genes are more highly expressed in lymphocytes than in other cell types (see Klein, 1986).

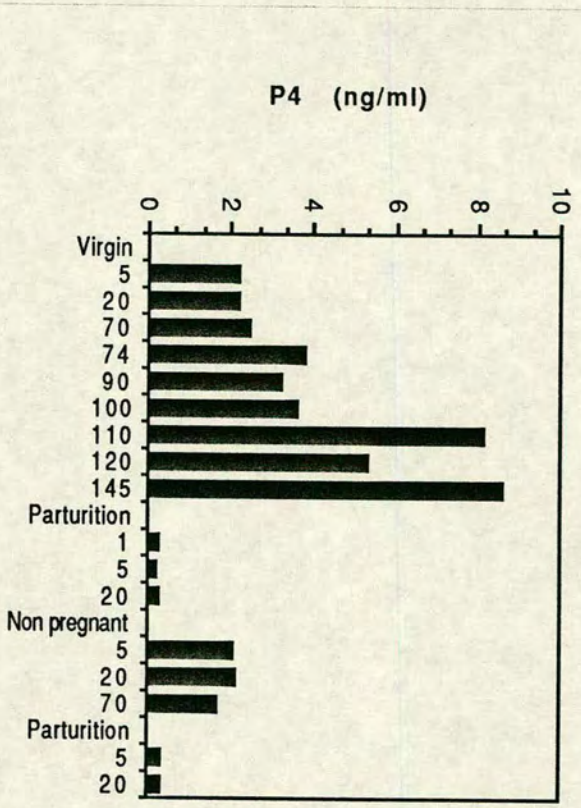
The milk protein genes are highly expressed mammary-specific genes, no signal being apparent at the level sensitivity of Northern blotting, in the other tissues. These nine probes were used in the sheep pregnancy and lactation time-course analysis. In order to show that each probe hybridises to a message of the correct size, an RNA gel was run, together with lambda/yeast RNA markers (from BRL). The gel was cut into strips and each strip was hybridised to one of the ten probes (including lambda DNA) (see figure 3.3). The bands obtained with each probe have been aligned in figure 3.3. BLG, α-lactalbumin and κ-casein gave similar-sized signals of about 800 nucleotides, whereas α_{s1}-, α_{s2}- and β-casein gave larger signals of about 1200-1400 nucleotides, again as expected. Similarly, BoLa (1700 nucleotides - see Ploegh et al., 1980), β-actin (about 2000 nucleotides - see Ponte et al., 1983) and rRNA (18S (2400 nucleotides) and 28S (6500 nucleotides) bands are seen - see Sollner-Webb and Reeder, 1979) signals were seen. Thus, each probe hybridises to RNA species of expected sizes.

Figure 3.3. Determination of RNA size. 2.0 μg of total mammary RNA from a 1-day lactating gimmer was run on a denaturing gel, together with lambda/yeast RNA markers (Pharmacia). The mammary and a single liver sample (1 day lactating gimmer) were probed with the nine probes described in the legend to figure 3.2. These are shown, together with marker sizes, to show that the hybridising signal is of expected size. The different strips show different exposures of a pre-flashed film placed at -70°C, in the presence of two intensifying screens. rDNA - 30 min. exposure; BLG, α_{s1}-, α_{s2}-, β- and κ-caseins - 16 hour exposure; α-lac, BoLal, β-actin - 70 hour exposure.

BLG
α-lactalbumin
αs1-casein
αs2-casein
β-casein
κ-casein
BoLa
rDNA
mam. liver β-actin
7.46/9.5
4.40
2.37
1.35
0.24

Figure 3.4. Serum progesterone and prolactin levels during pregnancy. P4 = progesterone, PRL = prolactin. From left to right: virgin, 5- 20-, 70-, 74-, 90-, 100-, 110-, 120- and 145-day pregnant gimmers, 0 days post-partum ("parturition"), 1-, 5- and 21-day lactating gimmers, nonpregnant draft, 5-, 20- and 70-day pregnant, 0 days post-partum draft ("parturition"), 5- and 20-day lactating draft. These assays were kindly performed by Dr. A. S. McNeilly (MRC Centre for Reproductive Biology) using the methods described by McNeilly and Andrews (1974) and McNeilly (1984).


PRL (ng/ml)
0
100
200
300
400
500
600
Virgin
5
20
70
74
90
100
110
120
145
Parturition
1
5
20
Non pregnant
5
20
70
Parturition
5
20


P4 (ng/ml)
0
2
4
6
8
10
Virgin
5
20
70
74
90
100
110
120
145
Parturition
1
5
20
Non pregnant
5
20
70
Parturition
5
20

3.2 SERUM HORMONE LEVELS DURING PREGNANCY AND LACTATION

20-30 mls of blood was taken from each ewe about an hour before its sacrifice and left to clot overnight. The serum was then removed leaving the clot and frozen at -20°C until assayed. Hormone assays for the major mammogenic hormone, progesterone and the major lactogenic hormone, prolactin, were kindly performed by Dr. A. McNeilly (MRC Centre for Reproductive Biology, Edinburgh). Figure 3.4 shows bar charts showing amounts of progesterone and prolactin present in sheep blood serum during pregnancy and early lactation.

These results showed that progesterone levels were undetectably low in the virgin and non-pregnant sheep. Progesterone levels rose immediately after pregnancy began and continued to rise throughout pregnancy, doubling between days 100-110 and reaching a peak just before parturition (day 145). Levels fell sharply at birth, being undetectable just after parturition and remained very low during lactation. Fewer samples were available late in draft ewes, so the rise in the level of progesterone was not seen. The rise in progesterone levels during pregnancy, particularly the sharp rise two-thirds of the way through the pregnancy and the sharp fall at parturition, has been described previously (see Heap and Flint, 1984).

Prolactin levels rose just after the start of pregnancy, then fell to almost undetectable levels for the greater part of the pregnancy. Placental lactogen levels have been found to rise during this time and then apparently carry out functions performed by prolactin. Levels of prolactin rose just before parturition and then remained at levels found in the virgin animal, rising to high levels by peak lactation (day 20). The draft ewes showed much higher levels of serum prolactin throughout pregnancy, levels being similar to the peak levels obtained in the first pregnancy

animals. During lactation prolactin levels were twice as high as in the first pregnancy animals. Progesterone levels were similar during both time-courses. It is not clear why there should be such a large difference between the gimmer and draft ewe prolactin levels. Seasonal differences in serum prolactin levels have been described previously; but these differences do not appear to affect lactation (Cowie, 1984).

It is, therefore, clear that progesterone levels rise during pregnancy, doubling at day 100-110 (just after secretory activity becomes detectable) and fall just before lactation. Prolactin levels are low (after an initial rise in levels early in pregnancy) during pregnancy and rise as progesterone levels fall at the end of pregnancy.

3.3 EXPRESSION OF BLG, α-LACTALBUMIN AND CASEIN GENES DURING PREGNANCY IN SHEEP.

2.0 μg of RNA from each time-point was run on formadehyde-MOPS agarose gels, transferred to nylon membranes and probed for the presence of each of the six milk protein gene mRNAs. Figure 3.5 shows the six Northerns and the three "control" Northerns.

Figure 3.5 also shows levels of β-actin and BoLa mRNA and rRNA levels (5, 10 and 2 μg of total RNA, respectively), throughout the time-course. These results show that some of the draft ewe RNAs are degraded. Nevertheless, these results show that the draft ewe time-course has similar regulation of milk protein gene expression as the gimmers.

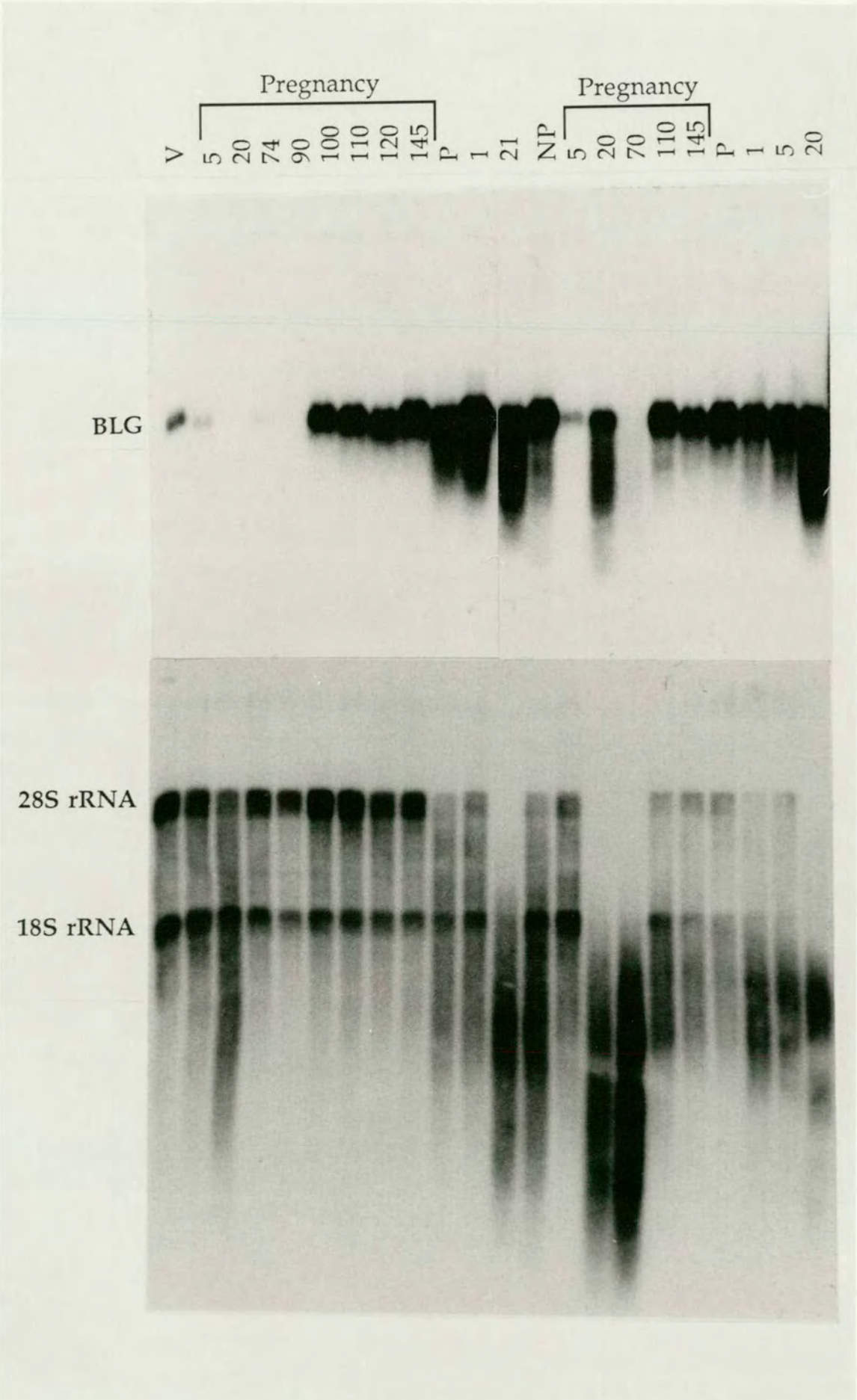
Interestingly, β-actin mRNA levels appear to decrease as pregnancy proceeds and during lactation. This may simply reflect a decline in its levels relative to total

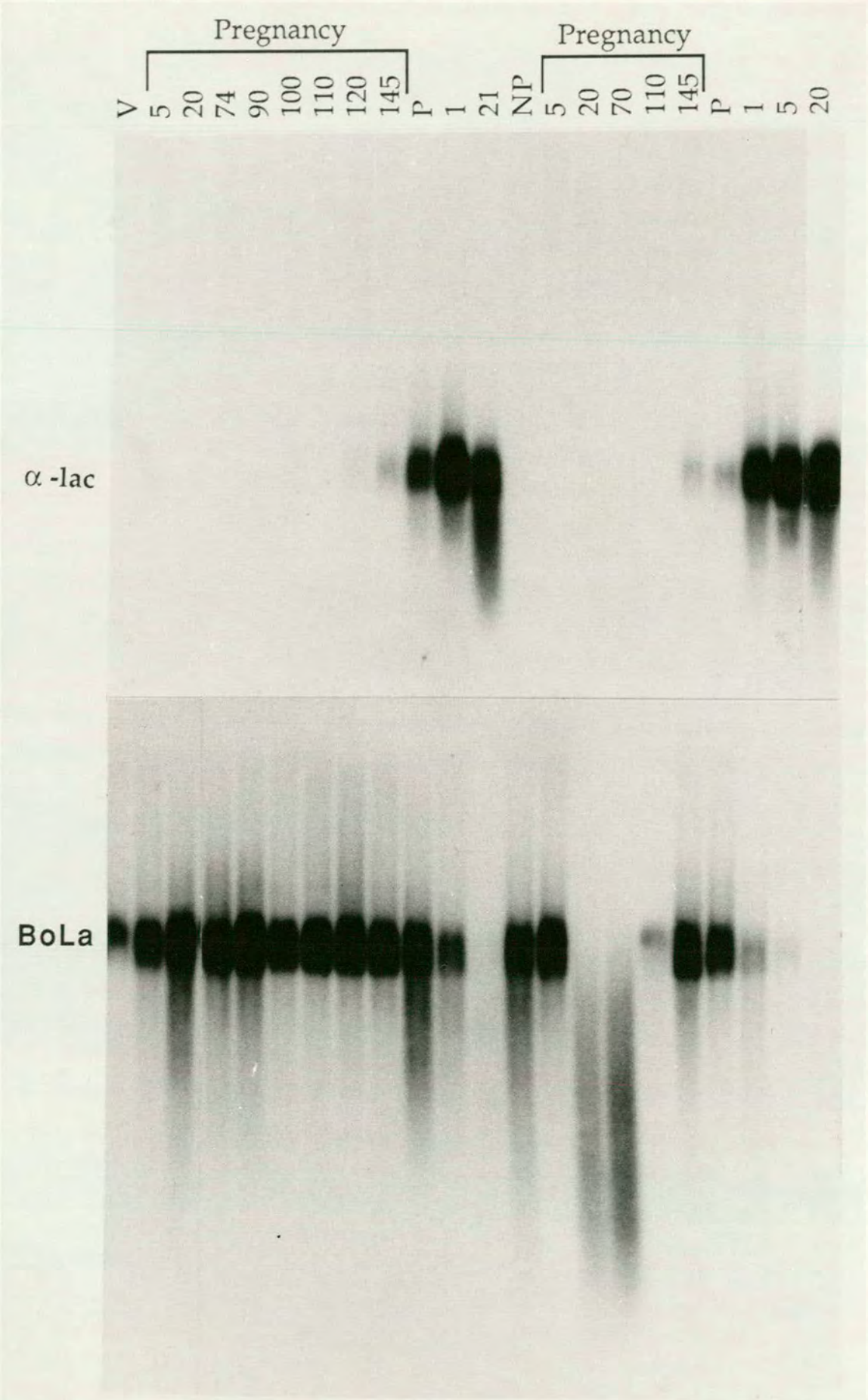
RNA levels, due to the increase in milk protein gene mRNA levels.

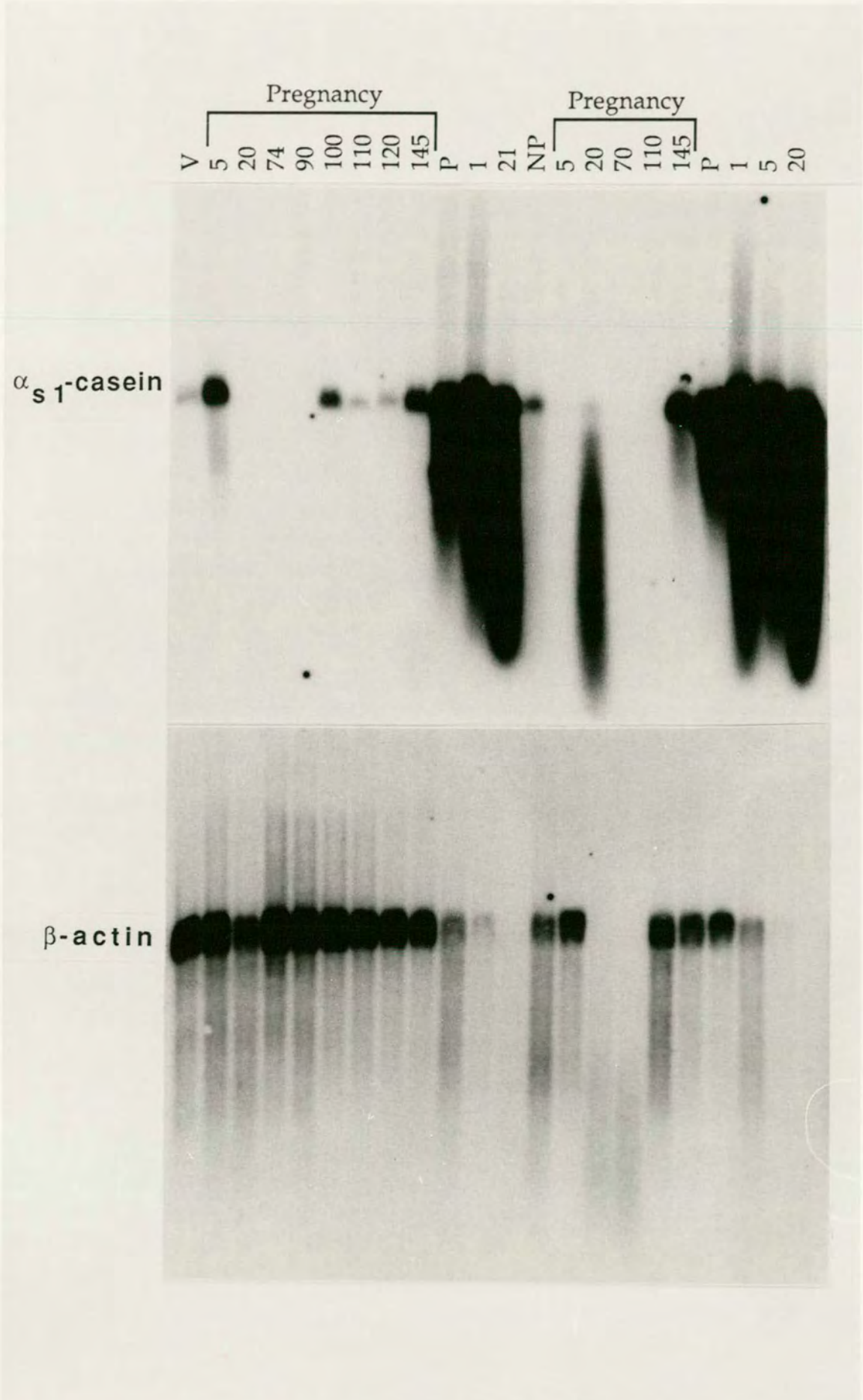
BLG mRNA (figure 3.5) was present in the virgin mammary gland. Levels fell just after the start of gestation, but rose sharply between days 90 and 100. mRNA levels continued to rise from day 100, reaching a peak at day 21 of lactation. Much higher BLG mRNA levels were present in the draft ewes throughout pregnancy than were seen in gimmers at these stages of pregnancy. BLG mRNA levels rose to day 21 of lactation.

The α-lactalbumin gene follows a different pattern of expression. Low levels of α-lactalbumin mRNA were present just after the start of pregnancy but levels declined quickly. α-lactalbumin mRNA became detectable again just before parturition (day 145) and rose sharply at parturition and during lactation. Little α-lactalbumin mRNA was seen in the non-pregnant draft ewe or at any point in the pregnancy, until day 145. Levels then rose to day 20 of lactation.

α_{s1}-casein mRNA levels increased in a similar manner to that of BLG, in both the first pregnancy and in draft animals; showing the same initial expression in the virgin and 5-day pregnant animals which fell at day 20, rising again from day 100 of pregnancy to day 21 of lactation. Levels of α_{s1}-casein mRNA, however, appear to be much lower between days 100-145, rising much more from day 145 of pregnancy, to lactation, than is observed for the BLG mRNA. β-casein mRNA levels follow a time-course of expression similar to that of the α_{s1}-casein gene.


Pregnancy
Pregnancy
V
5
20
74
90
100
110
120
145
P
1
21
NP
5
20
70
110
145
P
1
5
20
BLG
28S rRNA
18S rRNA


Pregnancy
Pregnancy
V
5
20
74
90
100
110
120
145
P
1
21
NP
5
20
70
110
145
P
1
5
20
α -lac
BoLa


Pregnancy
Pregnancy
V
5
20
74
90
100
110
120
145
P
1
21
NP
5
20
70
110
145
P
1
5
20
αs1-casein
β-actin

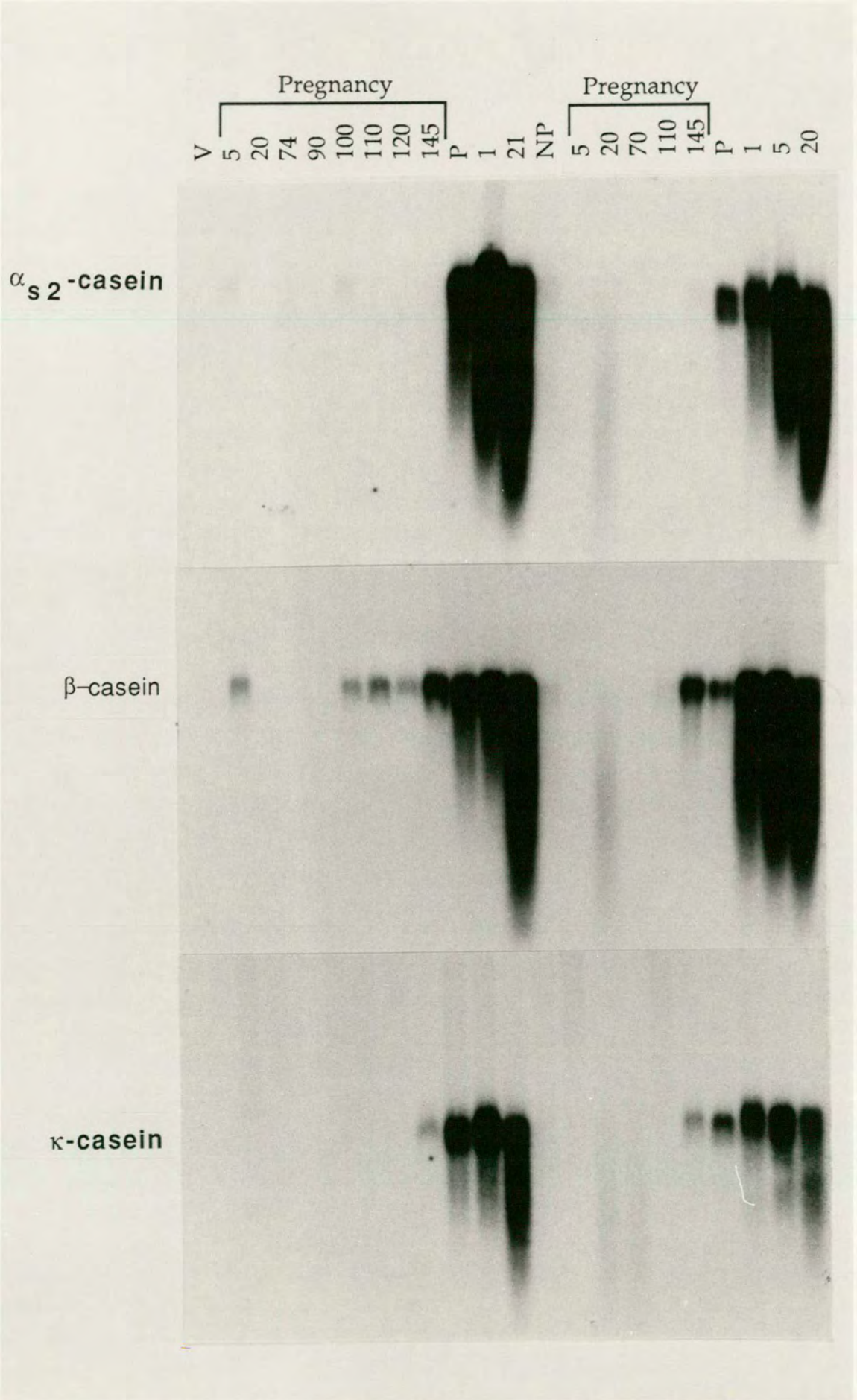
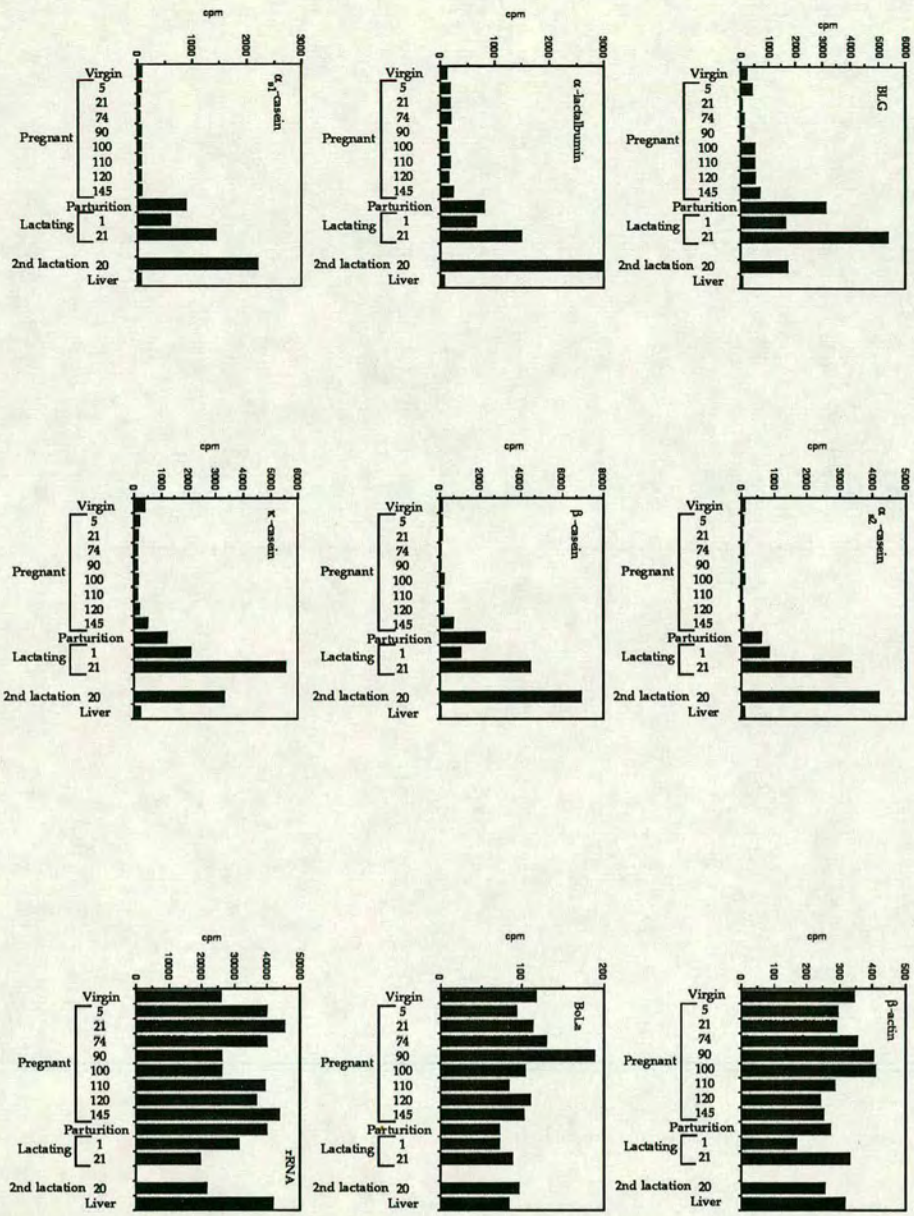

Pregnancy
Pregnancy
V
5
20
74
90
100
110
120
145
P
1
21
NP
5
20
70
110
145
P
1
5
20
α_{s2}-casein
β–casein
κ-casein

Figure 3.5. Milk protein mRNA levels during pregnancy. 2.0 μg of mammary RNA from the different time points durning pregnancy (see below) were probed for the presence of mRNAs for BLG, α-lac, a_{s1}-, α_{s2}-, β- and κ-caseins and of rRNA. 5.0 and 10.0 μg of total RNA was probed for the mRNAs for β-actin and BoLa, respectively (see legend to figure 3.2). From left to right: virgin (V), 5-, 20-, 74-, 90-, 100-, 110-, 120- and 145-day pregnant gimmers, 0 days post-partum (P), 1- and 21-day lactating gimmers, nonpregnant draft (NP), 5-, 20-, 70-, 110- and 145-day pregnant, 0 days post-partum draft (P), 1-, 5- and 20-day lactating draft.

Figure 3.6. Milk protein mRNA levels during pregnancy. 0.25 μg of total RNA from the gimmer pregnancy and lactational time points, 20-day lactating draft ("2nd lactation 20") and liver RNA from the 1-day lactating gimmer was slot-blotted and probed for the nine RNAs described in figure 3.5 (see Materials and Methods). The slot blots were cut and the amount of hybridisation determined by Cerenkov counting. Counts per minute (cpm) are shown.


cpm
Virgin
5
21
74
Pregnant
90
100
110
120
145
Parturition
Lactating
1
21
2nd lactation
20
Liver
αs1-casein
α-lactalbumin
BLG
κ-casein
β-casein
αs2-casein
rRNA
BoLa
β-actin

Little α_{s2}-casein mRNA was detectable until after parturition and then levels rose quickly. No κ-casein mRNA was detectable until day 145 of pregnancy in both gimmer and draft ewes. Levels of κ-casein mRNA then rose to day 21 of lactation.

It was not feasible to use biopsy techniques due to vivisection rules and the potential heterogeneity of sheep mammary gland tissue, which would have required multiple biopsy samples to be taken, at least at early stages of pregnancy. Thus, sacrifice of animals has meant that multiple animals were not used at any time-point, due to expense. However, much mammary tissue was frozen in liquid nitrogen and multiple RNA preparations were made, in order to average out the effects of heterogeneity in the mammary gland.

Slot blot analysis was performed using a different RNA preparation for each time-point, from that used in the Northerns presented above. 0.25 μg of total RNA was slot-blotted using a manifold device, as described in Materials and Methods. For this analysis only the gimmer ewe time-course has been used so far. 50-100 ng of each probe was oligo-labelled; this should be in excess of the RNA signal. After hybridisation, each slot was cut and counted to determine the amount of bound radioactivity. The results are shown in figure 3.6. It is not possible to quantitate mRNA levels to get amounts of each milk protein mRNA relative to other milk proteins because different sized probes were used. In addition, some probes contained poly(A) tails of different sizes and/or vector sequences. Quantitative comparisons for each mRNA across the time-course should be possible. Furthermore, it should be possible to standardise probe conditions in order to make comparisons between different milk protein mRNA levels. The bar charts in figure 3.6 show the counts per minute which hybridised to RNA from each time-point. In addition to the gimmer time-course RNAs, RNA from the 20-day lactating draft ewe and 1-day lactating gimmer liver RNA were used, for comparison. Very little signal was found in liver RNA when using any of the

six milk protein gene probes (also see figure 3.2). Expected magnitude of hybridisation was observed with β-actin, BoLa and rDNA probes.

This analysis differed very little from the northern data shown in figure 3.5. As seen in figure 3.5, BLG mRNA was present in the virgin and early pregnancy mammary gland, but quickly declined. Levels increased at day 100, rising to day 21 of lactation. α_{s1}- and β-casein mRNA levels also followed this pattern of expression, although α_{s1}-casein mRNA levels did not rise appreciably above background levels until day 145. Background hybridisation was considerably higher in the case of α-lactalbumin than for the other milk protein genes. This is likely to be due to the long poly(A) tail present in the ovine α-lactalbumin cDNA (J.-L. Vilotte, personal communication). This is also suggested by the smear seen in longer exposures of northerns.

mRNA levels at around day 20 of lactation in gimmer and draft ewes show considerable differences. Thus, BLG mRNA levels were 2-3 fold lower in the draft ewe than in the gimmer. κ-casein mRNA was almost half as abundant in the draft ewe as it was in the gimmer. All other milk protein mRNA levels were higher in the draft ewe than in the gimmer. All three controls showed similar levels of hybridisation in the gimmer and draft lactating ewes' RNA. The significance of these differences in mRNA levels is unclear but they appear to be real differences and are not due to loading differences.

Figure 3.7. *HpaII* methylation. 10 μg of liver and mammary DNA from five time points (from left to right: V = virgin, P = 120-days pregnant gimmer, L = 21-day lactating gimmer, NP = non-pregnant draft, L = 20-day lactating draft) was digested with *HpaII* (H) and *MspI* (M) and probed with the 4.4 kb SS12 *BamHI* fragment (see text). The numbers show lambda marker sizes, in kb. The heavy horizontal bars indicate fragments seen in mammary and liver *HpaII* digests but not in *MspI* digestions. *MspI* digestion bands are indicated by circles. The thin horizontal line indicates the *MspI* fragment seen in some but not all DNA samples (this may be a polymorphic site).

HpaII/MspI
SS1
H M Liver
H M Mam.
H M Liver
H M Mam.
H M Liver
H M Mam.
H M Liver
H M Mam.
H M Liver
H M Mam.
V
P
L
NP
L
21.3
5.15/4.97
4.27
3.54
2.10/1.90
1.58
1.38

Figure 3.8. *HhaI* methylation. 10 μg of liver and mammary DNA from five time points (from left to right: V = virgin, 120 = 120-days pregnant gimmer, L = 21-day lactating gimmer, NP = non-pregnant draft, L = 20-day lactating draft) was digested with *HhaI* and probed with the 4.4 kb SS12 *BamHI* fragment (see text). The numbers show lambda marker sizes, in kb. The small circles indicate the large fragments seen in liver DNA *HhaI* digests, the large circles show fragments seen in the mammary DNA digests.

HhaI
SS1
Liver V Mam.
Liver 120 Mam.
Liver L Mam.
Liver NP Mam.
Liver L Mam.
21.3
5.15/4.97
4.27
3.54
2.10/1.90
1.58
1.38

3.4 METHYLATION OF THE BLG GENE IN SHEEP

Studies analysing chromatin structure and DNA composition have shown that there are often differences between the chromatin at, or near, expressing genes and non-expressing genes, or intergenic regions (see chapter 4 for discussion). Work describing base composition and particularly, DNA methylation, has also shown differences between expressed and non-expressed regions. The vertebrate genome is A+T-rich, 60% of the bases being adenines and thymines. CpG dinucleotides are present at 20% of their expected frequency in vertebrate DNA. Furthermore, 60-90% of all CpG dinucleotides in vertebrate DNA are methylated at the cytosine. However, there are regions which differ considerably from this mean. In particular, short G+C-rich regions of 1-2 kb are present. In these, the G+C content can be as high as 65% (compared to 40% for bulk DNA). The presence of these high G+C "islands" is also correlated with enrichment for the dinucleotide CpG and a lack of CpG methylation. In regions where CpGs are methylated there is an overall deficiency of the dinucleotide, whereas in G+C-rich regions near normal levels of CpG and lack of methylation are found. The reason for this appears to be that 5-methyl-cytosine is highly mutable to give thymine. Thus CpG deficiency is found to correlate with TpG and CpA excess. The presence of G+C-rich islands is at least partly due to the normal levels of CpG. The presence of such "CpG islands" is, therefore, predicted by a G+C content of over 50%, CpG levels approximately equal to GpC levels (and at levels expected from base composition) and undermethylation (see Razin and Cedar, 1984; Bird, 1986; Bird, 1987).

The presence of such CpG islands has been noted at the 5' ends of many genes, suggesting a possible role in the control of gene expression. All sequenced RNA polymerase II-transcribed "housekeeping" genes have CpG islands at their 5' ends,

although most have not been tested for hypomethylation. Some tissue-specific genes also have such islands (such as α(2)I-collagen, retinol-binding protein, α-globin, Thy-I and Class II histocompatibility genes, although many others show no evidence of CpG islands (for example, the growth hormone, fibrinogen and myoglobin genes) (see Bird, 1986).

Experiments in which transfected genes have been methylated (or not) show that methylation of the CpG island reduces transcription relative to the unmethylated gene (Razin and Cedar, 1984; Bird, 1986). Use of the DNA methylation inhibitor 5-azacytidine, can cause reactivation of expression (see Bird, 1986). Also, in tissue-specific genes CpG methylation is seen in tissues in which the gene is not expressed and undermethylation in tissue(s) in which it is expressed (e.g. chicken ovalbumin gene - Mandel and Chambon, 1979). It has also been noted that some genes show loss of CpG methylation on induction (see Razin and Riggs, 1980 for review). The rat γ casein gene shows undermethylation of some CpGs in the mammary gland of lactating females, relative to the liver (Johnson et al., 1983).

However, it is not clear whether methylation changes and differences are due to active expression of a gene; or whether loss of methylation is preliminary to transcriptional activation (see reviews sited above). Furthermore, current models of the evolution of such islands propose that they may have evolved due to increased methylation in vertebrates, compared to non-vertebrate ancestors and subsequent reduction in CpG levels by mutation; whilst promoters, particularly of housekeeping genes did not become methylated, perhaps due to the influence of bound factors. In such models a function for such islands is not necessary. Function(s) may, however, have arisen once these islands had evolved. Recent evidence suggests that at least one housekeeping gene transcriptional activator, Sp1, can bind methylated DNA, as well as unmethylated DNA *in vitro* (Holler et al., 1988; and references therein). Holler et al.

(1988) also showed that methylation did not inhibit transcription *in vitro* or *in vivo*. This indicates that CpG methylation may not be important for transcription. Thus, undermethylation may be a consequence of bound factors.

Whether it is a cause or consequence of gene expression, it is clear that some genes are undermethylated in expressing tissues, relative to non-expressing tissues (van der Ploeg and Flavell, 1980). Their work suggested that some loss of methylation at the β-globin gene locus occurs in expressing tissues. In the genes so far examined not all CpGs are undermethylated in expressing cells even within CpG-rich domains (for example see van der Ploeg and Flavell, 1980; Johnson et al., 1983). Methylation analysis is usually carried out using restriction enzymes which cleave at a sequence containing a CpG dinucleotide and which do not cleave if the CpG is methylated. The most commonly used enzymes are *HhaI* (GCGC) and *HpaII* (CCGG). *MspI* also cuts at CCGG but will cleave methylated or unmethylated DNA. It is therefore, useful for comparative digestions with *HpaII*.

Some preliminary digestions have been done using liver and mammary DNA from the time-course sheep. Virgin, 120-day pregnant and lactating gimmer ewe and non-pregnant and lactating draft ewe genomic DNAs were digested with *HhaI*, *HpaII* and *MspI*, run on 1% agarose gels and Southern blotted. These digests were probed with a 4.4 kb *BamHI* fragment derived from SS12 (see chapters 4 and 5), which contains about 2 kb of 5' flanking sequences and exon I to intron 3 of the ovine BLG gene. The results are shown in figures 3.7 and 3.8.

Figure 3.7 shows *HpaII* and *MspI* digests of the 10 genomic DNAs. Two fragments, both less than 1 kb in size were seen in the *MspI* digests and in *HpaII* digests of one of the phages encoding the ovine BLG gene, SS1 (see chapter 4). These fragments were not present in the mammary or liver *HpaII* digests, suggesting that the restriction sites are methylated. A third fragment, also less than 1 kb in size, was

present in two sheep DNAs and may represent a polymorphic *HpaII* site. These fragments were not present the mammary and liver *HpaII* digests, although some low levels of hybridisation were seen. Undermethylation in mammary DNA digests, relative to liver DNA digests, was not apparent. Two fragments, larger than 1.0 kb, but smaller than 1.38 kb were just visible in the mammary and liver DNA *HpaII* digests, but were absent from *MspI* digests. Other fragments are expected to be present in *MspI* digests, suggesting that they may not have resolved on this 1% agarose gel. It is, therefore, possible that other *HpaII* restriction sites are undermethylated in the mammary gland, relative to the liver.

HhaI digests are shown in figure 3.8. Again, *HhaI* digested SS1 DNA was used as control and the SS12 4.4 kb *BamHI* fragment was used as a probe. The DNA sequence of SS1, presented in chapter 4, shows that fragments of about 670 bp (*HhaI* site in intron 1 (+390 - two overlapping sites) to intron 2 (+1065)), 1400 bp (+1065 to intron 3(+2440)) and 340 bp (+2440 to +2778 (just 3' of the *BamHI* site)) should be seen. In addition a fragment larger than 1200 bp (from +390 to sequences 5' of -800) and one, or more, other fragments, should be seen (for more details see chapter 4, figures 4.3 and 4.4). The SS1 digest showed that a 3000 bp fragment is present, indicating that there are no *HhaI* sites in the sequences 3' of the *BamHI* site, until intron 1. In figure 3.8 the expected 340 bp fragment was not resolved. All larger fragments were seen. The 3000 bp, 1400 bp and 670 bp fragments are present in mammary DNA *HhaI* digests. The two smaller fragments (670 and 1400 bp) are apparently absent (or present at low abundance) and the largest fragment is present at much lower intensity, in liver DNA *HhaI* digests. Furthermore, liver DNA contains three fragments between 10-20 kb in size, which are present in the mammary DNA at much lower intensity. Thus, mammary DNA is apparently undermethylated at these sites, relative to liver DNA. In addition, (despite slightly greater loading) *HhaI* digest

fragments are present at a lower intensity in the virgin mammary DNA than in the 120-day pregnant mammary DNA digest. This may indicate demethylation of the BLG gene during pregnancy. Alternatively, it may simply indicate changes in cell population from the virgin mammary gland when much adipose tissue is present in the mammary gland, to more mammary epithelial cells during pregnancy and lactation. If the latter is correct, then adipose cells (and other cell types replaced by increases in mammary epithelial cell numbers) also show greater methylation relative to mammary DNA. In any case the undermethylated state of expressing tissue (mammary gland) relative to non-expressing tissue (liver) DNA seems clear. As mentioned above, similar methylation differences in the rat lactating mammary gland γ-casein gene were shown by Johnson et al. (1983). The results presented here suggest that not only is lactating mammary DNA undermethylated but virgin mammary DNA is also undermethylated. The possibility of changes in methylation during the life cycle has also been raised.

In contrast to the results of *HhaI* digestion, *HpaII* digests show no apparent methylation differences and show that mammary DNA may be methylated. From DNA sequence information 16 HpaII fragments are expected, most less than 100 bp, only three more than 300 bp in length and none larger than 600 bp. In addition, the sequence of about 1 kb of 5' flanking sequences in the probe, is not known. The SS1 digest shows, however, that all fragments are much smaller than 1.38 kb. Only two fragments were visualised. Thus, the full nature of the *HpaII* site methylation status remains to be determined.

3.5 DISCUSSION

Histological analysis and serum progesterone and prolactin levels show changes during pregnancy and lactation which have been described previously (see Denamur, 1974; Anderson, 1975; Heap and Flint, 1984; Cowie, 1984). Previous work (described in the introduction to this chapter) and this time-course (figure 4.1), have shown that secretory activity begins by day 90 of gestation, becoming extensive by day 100 (M. McClenaghan's unpublished results). Furthermore, hormone assays have shown that levels of prolactin rise at day 145 of gestation, whilst progesterone levels become undetectable at parturition. Denamur (1974) and Anderson (1975) have shown that RNA levels rise dramatically at parturition, in sheep. These findings suggest that days 90-100 and 145 of pregnancy, parturition and early lactation are key time-points in the regulation of lactogenesis. The onset of secretory activity (lactogenesis) does not follow, nor is accompanied by, changes in levels of the major mammogenic (progesterone) and lactogenic (prolactin) hormones. Thus changes in absolute levels of these hormones do not appear to be responsible for the lactogenesis. Furthermore, progesterone levels rose from the start of pregnancy, doubling between days 100-110 of gestation. Similar increases have been described in sheep and in many other mammals, including guinea pig, man and rat. Sheep, man and rat progesterone levels rise sharply from mid-pregnancy, coinciding with initiation of secretory activity. This is also seen in the gimmer time-course presented here (M. McClenaghan's results).

Total mammary RNA from each time-point has been analysed on Northerns and by slot-blotting and the temporal regulation of all major ovine milk protein genes has been investigated. Although these techniques have not yet been used to directly compare levels of each mRNA present during pregnancy and lactation, it has, nevertheless, been possible to demonstrate coordinate and non-coordinate regulation of

milk protein gene expression. BLG, α_{s1}- and β-casein genes show similar regulation, the levels of each mRNA changing at the same time-points during pregnancy. Analysis of a time-course in transgenic mice containing the ovine BLG gene showed a similar patterns of expression for the ovine BLG and the mouse β-casein genes, whilst the mouse WAP gene followed a different pattern of expression (S. Harris, unpublished results). This suggests that the 16.2 kb BLG gene construct contains the sequences required for correct temporal regulation.

α-lactalbumin and κ-casein mRNA levels rise at day 145. This may correlate with increases in levels of prolactin at this time (see figure 3.4).

α_{s2}-casein gene does not appear to be expressed prior to parturition. At parturition, mRNA levels rise quickly. All six milk protein mRNAs show rapid increases at parturition. These continue to rise to day 20 of lactation. This rise in mRNA levels may reflect a stimulatory effect of prolactin and/or fall in progesterone levels. α_{s2}-casein gene expression may be particularly sensitive to progesterone levels, as no expression was seen prior to parturition.

It was noted in figure 3.5 that β-actin mRNA levels fall through pregnancy and into lactation. This is again apparent from the slot-blotting data in figure 3.6. β-actin mRNA levels decreased through pregnancy and lactation. A similar decrease in mRNA levels was seen for BoLa. This "reduction" in levels of β-actin and BoLa mRNAs is probably due to the increase in levels of milk protein mRNAs which causes an apparent decrease in β-actin and BoLa mRNAs, relative to total RNA, but probably does not reflect changes in their gene expression. Thus, absolute levels of their mRNAs probably remain quite constant. Milk protein mRNA levels rise from low amounts prior to day 100 of gestation (and more particularly, day 145 of gestation) (this time course) to make up 60-80% of total poly $(A)^+$ mRNA during lactation (see Mercier and Gaye, 1983).

This analysis of milk protein gene expression has shown that the BLG gene is expressed in the virgin mammary gland. Coordinate regulation of BLG mRNA levels with other milk protein mRNAs has been shown. BLG mRNA levels increase when secretory activity is first observed. Its temporal regulation, like that of the other milk protein genes is regulated during pregnancy. It is not yet known whether this increase in BLG mRNA levels is transcriptionally or post-transcriptionally regulated. Evidence for post-transcriptional control of rabbit and rat milk protein genes has been discussed in chapter 1. Large amounts of sheep mammary tissue were frozen under liquid nitrogen and will be used for transcriptional run-on experiments. These may indicate whether increase in BLG (and other ovine milk protein) mRNAs is due to an increase in transcription. Initial experiments have been attempted and difficulties in isolating sufficient nuclei have been encountered (C. B. A. Whitelaw and myself). This is not a problem during late pregnancy stages but for early pregnancy time-points large amounts of mammary tissue will need to be used to obtain sufficient nuclei for run-on experiments.

The RNA data showed that the BLG gene is expressed in the virgin mammary gland. In the past, work describing DNA methylation of vertebrate genes has shown that many genes are undermethylated in expressing, relative to non-expressing, tissues. Mapping methylation sensitivity of restriction enzymes can, therefore, be indicative of transcriptionally active genes. Initial results indicate that the ovine BLG gene is active in the virgin mammary gland, in agreement with the RNA data. It is possible that the BLG gene (and other milk protein genes) are induced at puberty or before birth. As stated in chapter 1, the secretion of a milk-like substance by the newborn has been demonstrated in man (see Hiba et al., 1977) and in other mammals (see Knight and Peaker, 1982). It is thought that this secretion occurs as a result of the influence of maternal and/or placental mammogenic and lactogenic hormones. It is therefore,

possible that milk protein genes become primed at this stage and low level expression is maintained, at least in females. It may be possible to examine mammary DNA methylation status from fetal, pre- and post-pubertal stages to investigate the stage at which these genes become transcriptionally active. DNaseI hypersensitivity sites also show tissue-specificity in some genes. Investigation of changes in DNaseI hypersensitive sites can also be indicative of gene activation (see chapter 4 for more details). Furthermore, methylation analysis and DNaseI hypersensitivity may map domains which are important for transcriptional control. A fuller discussion will be found in chapter 4.

The large increases in milk protein mRNA levels during pregnancy remain to be investigated. The apparently "correct" tissue- and temporal-regulation of the ovine BLG gene in transgenic mice and abundant expression (Simons et al., 1987; S. Harris, unpublished results) indicate that control of BLG gene expression can be analysed in the mouse model system.

Chapter 4. CHARACTERISATION AND SEQUENCING OF THE GENE ENCODING OVINE β-LACTOGLOBULIN

4.1 INTRODUCTION

In chapter 3 I showed that the ovine BLG gene is tissue-specifically and temporally regulated. Cloning and sequencing of the gene are prerequisite to its dissection for control of expression. DNA sequencing of a gene provides information about its structure and also allows a search for sequences which could potentially control its expression. In particular, the exonic structure of a gene is informative about its evolution and the structure and function of the protein product it encodes (see chapter 6).

Control of gene expression in eukaryotes involves *cis*-acting DNA sequences and in a number of genes these have been dissected by deletional and mutational analysis. These sequences are short, 10-12 bp motifs which are recognised by transcription factors. Two types of DNA sequences are important for transcriptional regulation, promoters and enhancers. Promoter sequences are required for accurate and efficient transcription initiation. They comprise the proximal elements (the TATA box and sequences around the transcription initiation site (Breathnach and Chambon, 1981)) and the distal "upstream promoter elements" (UPEs). The TATA element is present in most RNA polymerase II-transcribed genes. Its deletion leads to lowered transcriptional efficiency and heterogeneous transcriptional start sites (Mathis and Chambon, 1981). Transcriptional rates are controlled by the presence of one or more UPEs. The strength of a promoter will depend on the number and type of elements. They are typically present in the -40 to -110 region. Some UPEs are

found in many promoters, for example the CCAAT box element (Breathnach and Chambon, 1981); others may be gene-specific. UPEs can act independently of orientation but are largely position-dependent. Changes in nucleotide number between them and the TATA element can lead to a reduction in the level of transcription (see Maniatis et al., 1987, for a recent review).

Enhancers can regulate transcription in either orientation and can act at considerable distances, relative to cis-linked promoters. They can function either upstream or downstream of a transcription unit. A number of genes have been shown to contain transcriptional control sequences in the 3' flanking region (for example, an enhancer is present in the 3' flanking sequences of the human β-globin gene (Kollias et al., 1987)) and some have control elements within introns (for example, the immunoglobulin heavy chain gene enhancer (Gillies et al., 1983; Banerji et al., 1983)). Short sequence motifs in enhancers have been recognised as being important and are sequences to which regulatory proteins bind. Some of these motifs are also found in UPEs .

The sequence elements are often symmetrical and may be repeated several times. Some motifs are present in many enhancers, for example, the SV40 and polyoma enhancer "core" sequence is also present in the immunoglobulin heavy chain gene enhancer (Gillies et al., 1983; Banerji et al., 1983). The core sequences may be recognised by the same regulatory protein or may form part of the recognition sequences for different proteins (e.g. a number of proteins that bind to the CCAAT box-like sequences have been characterised - see Dorn et al., 1987; Chodosh et al., 1988; Santoro et al., 1988; and references therein). Other regulatory sequences are sometimes found repeated three or four times, for example the heat-shock element in many of the heat shock promoters (see Pelham, 1985); the number of heat shock-responsive elements present has an additive effect on heat shock induction

(Dudler and Travis, 1984). Similar findings have been obtained for the heavy metal-responsive element in the mouse metallothionein-I promoter (Stuart et al., 1985) and the inducible element in the human β-interferon gene (Goodburn et al., 1985). In this respect these sequences can act in a similar manner to enhancers (Bienz and Pelham, 1986).

Initiation of most RNA polymerase II-transcribed genes requires the presence of a TATA box element. A protein factor, TFIID (also known as BTF1) binds to this element. This is among the first steps leading to transcription initiation. A transcription complex is formed, with RNA polymerase II and other general transcription factors (Zheng et al., 1987; and references therein). It is not clear what role factors binding to UPEs and/or enhancer elements play in initiation complex formation. It is also not clear whether transcriptional elements interact with TFIID, RNA polymerase II, the other general transcription factors, or some or all of these. It has recently been shown that several transcription factors can interact directly with TFIID (Horikoshi et al., 1988; and references therein). The possibility that some transcriptional regulators may interact with RNA polymerase II has been indicated by the cloning of yeast, Drosophila and mouse RNA polymerase II genes (Allison et al., 1985; Corden et al., 1985). The predicted amino-acid sequence shows that a unit of seven amino-acids (at the carboxy end) is repeated 26 times in the yeast and 52 times in the mouse RNA polymerase II polypeptides. This unusual structure may interact with activation regions of transcriptional regulators, thereby influencing preinitiation complex formation (Sigler, 1988).

Promoter and enhancer elements are involved in temporal and tissue-specificity control of gene expression. A number of different mechanisms have been shown to offer such control. One of the simplest methods may involve the *de novo* synthesis of a transcription factor (no clear examples of this have emerged,

although *c-fos* may be an example of this - its expression is induced by many serum growth factors, phorbol esters and cAMP (e.g. see Gilman (1988)); and it is now emerging that *c-fos* is probably a transcription factor (Vogt et al., 1987; Lech et al., 1988; Landschulz et al., 1988)). Interactions with other proteins seems to provide an important mechanism for activation or inactivation of transcription factors. The human glucocorticoid receptor is apparently present in an inactive form, bound to a heat shock protein (*hsp90*) in the cytosol (see Pratt et al., 1988; and references therein). On induction with glucocorticoids it is freed from hsp90 and enters the nucleus to interact specifically with glucocorticoid response elements. Some other transcription factors consist of two subunits. The CCAAT-binding family of proteins appear to have a polypeptide which contains the DNA-binding domain but which appears to bind DNA only in association with a second subunit (Chodosh et al., 1988; Hatamochi et al., 1988). Landschulz et al. (1988) have postulated that some factors form a "leucine zipper" in which two monomers dimerise by interactions between a set of leucine residues. This "zipper" is predicted from the amino-acid sequences of the factors AP1, c-fos, c-myc and GCN4. It may also allow heteromers to associate. Yet another mechanism of activation of transcription factors is by post-translational modification, such as phosphorylation. Sorger and Pelham (1988) have recently cloned the yeast heat shock factor and have shown that it exhibits temperature-dependent phosphorylation. A cAMP responsive element is present in a number of genes. Cyclic AMP presumably acts through the action of a kinase on one or more transcription factors (Jameson et al., 1987). The transcription factor, AP1 appears to be induced by phorbol esters which affect protein kinase C (Angel et al., 1987).

A further form of control by different interactions between transcription factors has involved displacement. A factor may bind to the DNA preventing binding

by another factor, thereby perhaps repressing transcription. The progesterone and glucocorticoid receptors appear to bind very similar, or overlapping, recognition sequences (von der Ahe et al., 1985, 1986; Strahle et al., 1987). The antagonistic actions of glucocorticoids and progesterone have been described (for example, see Ganguly et al., 1982). Thus, exclusion of one factor by the other could control expression. Another example of this was described recently (Barberis et al., 1987). A factor binds to sequences overlapping with the proximal CCAAT-binding element of the sea-urchin histone H2B gene in the embryo. Here the gene is not expressed. In expressing cells, spermatocytes, a CCAAT-factor binds and expression is seen. This would appear to be a clear example of exclusion of one factor by binding of another.

The ovine β-lactoglobulin (BLG) gene is a mammary gland-specific gene which is expressed during pregnancy and lactation. Its expression is thought to be under the control of a number of peptide and steroid hormones, such as prolactin, progesterone, oestrogen and glucocorticoids. Analysis of the DNA sequence of the gene may provide evidence for the possible location of sequences controlling expression. For example, oestrogen, progesterone and glucocorticoid-responsive elements have been mapped using mutational analysis and using DNaseI footprinting experiments (for example, see Scheidereit et al., 1986). These sequences can act as enhancer elements. The presence of these sequence motifs in the BLG gene sequence would be evidence for possible control of BLG gene expression by these hormones.

Comparison of the sequences of the BLG gene with other milk protein gene sequences can also be carried out. The presence of similar sequences in otherwise unrelated genes may indicate sequences important for control of expression by the mammary gland. Hall et al. (1987) have identified a relatively large region of about 32 bp which appears to be present in the 5' flanking region of some milk protein

genes of cow, guinea pig, man and rat, between -140 and -110. This sequence may be involved in tissue-specificity or hormonal regulation of milk protein genes.

The presence of any of these sequence motifs in the BLG gene does not mean that these sequences are required for expression. Mutational studies will be required to show if any of these sequences are actually involved in the control of expression. Recent evidence has also shown that some proteins can bind to different DNA sequence motifs (Davidson et al., 1988); the progesterone and glucocorticoid receptors recognise similar sequence motifs. Thus it can be difficult to judge the value of sequence elements found by DNA comparisons.

In this chapter I present data on characterisation of the genomic clones encoding ovine BLG gene and present the DNA sequence together with analysis of this sequence. I show that SS1 contains the sequences which are known to be absolutely required for a gene to be functional and discuss results which suggest that it encodes a functional ovine BLG gene. Evidence for the presence of putative transcriptional control elements is presented. The initial work was done by A. J. Clark and is described below.

High molecular weight genomic DNA prepared from sheep spleen was partially digested with *Sau3A* and size fractionated DNA fragments of 14-20 kb were ligated into the *BamHI* sites of the lambda phage, EMBL3 (Frischauf et al., 1983), packaged and plated. The library was screened using a plasmid containing a cDNA for ovine BLG (p931). p931 contains a cDNA insert of about 500 bp, excluding the polyA tail (Gaye et al., 1986). Six positively hybridising plaques were isolated and DNA was prepared from four of them. The four clones were restriction mapped using six restriction enzymes and southern blotting and hybridisation, with p931 and specific 5' and 3' fragments derived from p931, was carried out (A. J. Clark's results; see Ali and Clark, 1988). Figure 4.1 shows restriction maps of the four

clones. The phages contain inserts of between 14 and 17.5 kb and overlap over the greater part of their lengths. Examination of the restriction maps shows that the four clones are very similar, although a number of differences are seen. Clones SS2 and SS12 have a *HindIII* site 1.9 kb from the 3' end of the BLG gene. This site is absent from SS1 and SS11. An *SphI* site present near the 5' end of the gene is absent from SS12. Three other restriction site differences are present at the boundaries of the clones SS1, SS2 and SS12.

The four clones described above are very similar, suggesting that they may indeed encode the ovine BLG gene. SS1 and SS11 appear to be identical, as do SS2 and SS12, indicating that SS1/SS11 and SS2/SS12 may represent two alleles of the ovine BLG gene (see chapter 5). Southern blotting of sheep genomic DNA restriction enzyme digests and hybridisation with p931 gave bands whose presence was predicted from the restriction maps of the clones (see section 4.3). Finally, the entire SS1 insert (the 16.2 kb *SalI* fragment), the 10.5 kb SS1 *SalI/XbaI* and the 10.85 kb SS12 *SalI/XbaI* fragments have been injected into mouse embryos to create transgenic mice. All three constructs work in transgenic mice to produce a protein which appears to be identical to ovine BLG, when examined by SDS/polyacrylamide gel electrophoresis and Western blotting techniques (Simons et al., 1987; chapter 5). SS1 and SS12 therefore, encode functional copies of the ovine BLG gene.

Furthermore, the initial DNA sequencing of SS1 was done by A. J. Clark. 5' flanking, exon I and exon II sequences were determined. I have confirmed and extended this sequence (presented below) to obtain the entire sequence from -40 to the *XbaI* site (+6578).

Figure 4.1 Restriction maps of BLG clones. The clones are aligned using common restriction enzyme sites. The extent of hybridisation of the BLG cDNA p931 (Gaye et al., 1986) and specific 5' and 3' *PstI* fragments isolated from the cDNA, is shown above the clones as horizontal lines. L and R refer to the left and right phage arms, respectively. These maps are shown courtesy of A. J. Clark.

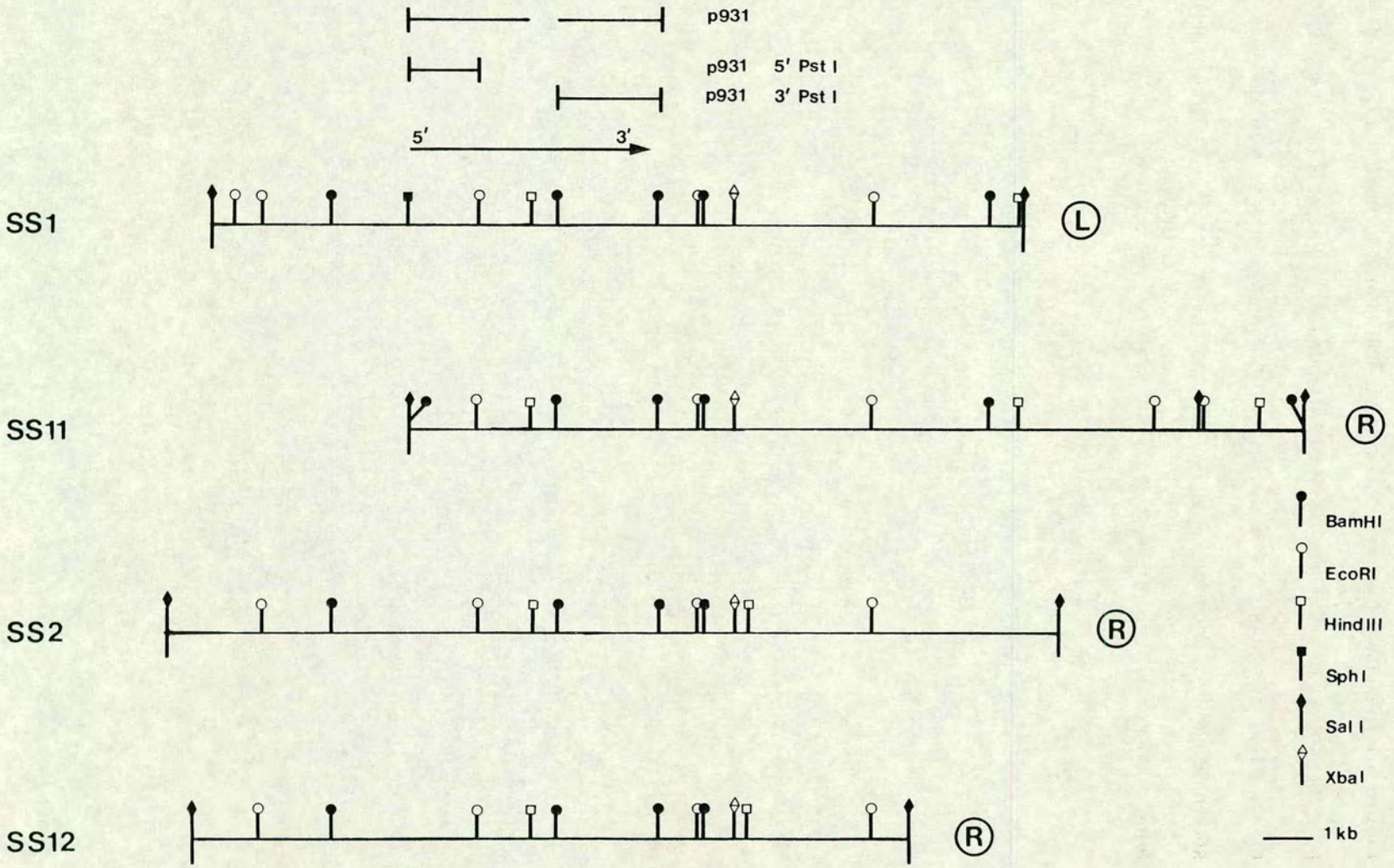
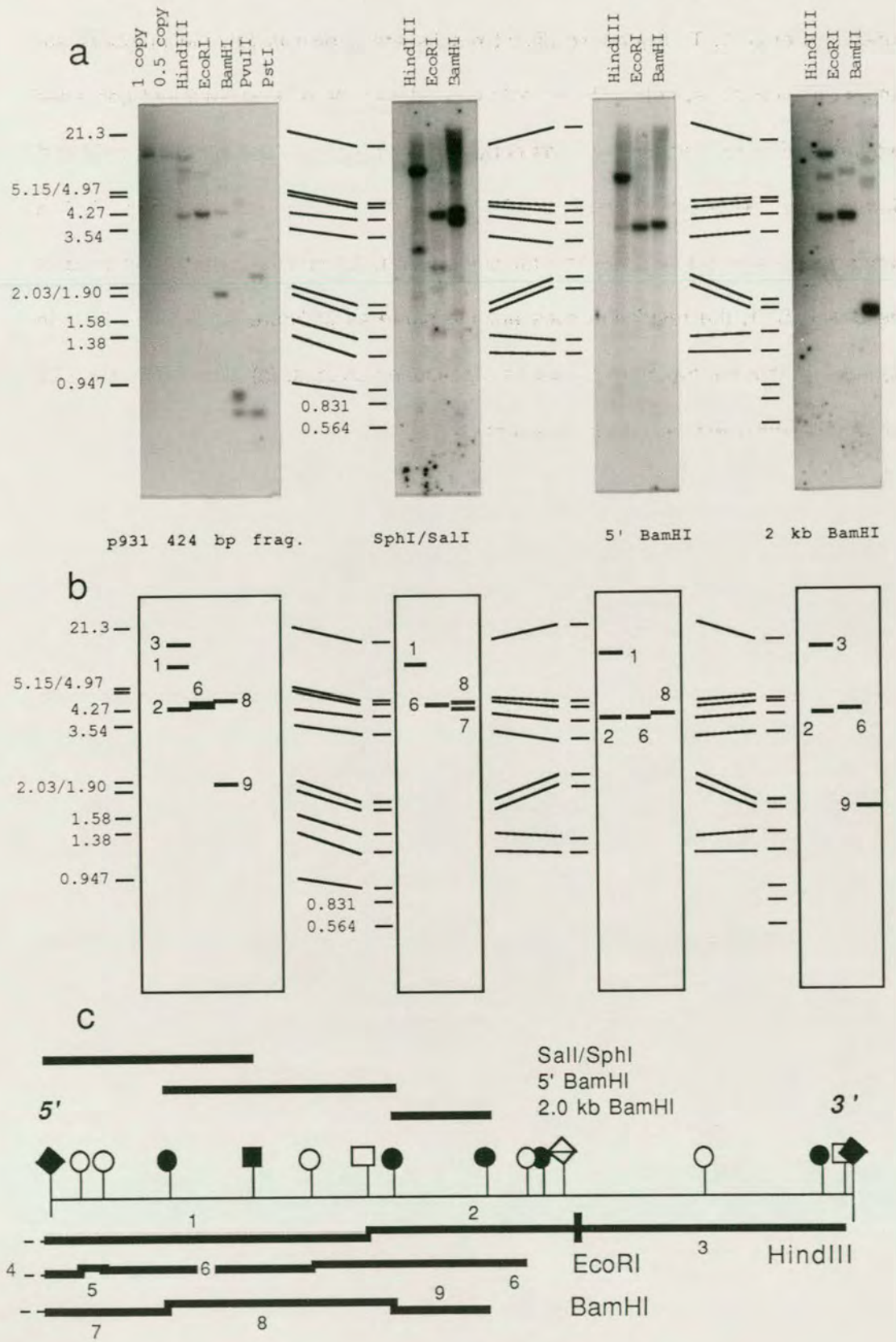

p931
p931 5′ Pst I
p931 3′ Pst I
5′
3′
SS1
L
SS11
R
SS2
R
SS12
R
BamHI
EcoRI
HindIII
SphI
SalI
XbaI
1 kb

Figure 4.2 Genomic Southern mapping of the BLG gene. 10 μg of restricted sheep genomic DNA was probed with the 424 bp *PstI* fragment of p931 and with three subclones of SS1. The genomic digest results are shown in (a), the numbers are lambda marker sizes, in kb. The digests probed with the 424 bp fragment contained two lanes of *HindIII*-digested SS, as copy control. (b) shows the *HindIII*, *EcoRI* and *BamHI* bands whose presence is predicted from the phage maps, shown in (c). The bands in (b) and (c) are labelled for comparison. (c) shows a restriction map of the phage SS1, the restriction sites are the same as in figure 4.1. Dotted lines in (c) indicate that the fragment continues 5' to cloned phages. (c) also shows the SS1 subclones which were used as probes in (a).


a
1 copy
0.5 copy
HindIII
EcoRI
BamHI
PvuII
PstI
HindIII
EcoRI
BamHI
HindIII
EcoRI
BamHI
HindIII
EcoRI
BamHI
21.3
5.15/4.97
4.27
3.54
2.03/1.90
1.58
1.38
0.947
0.831
0.564
p931 424 bp frag.
SphI/SalI
5' BamHI
2 kb BamHI
b
21.3
5.15/4.97
4.27
3.54
2.03/1.90
1.58
1.38
0.947
0.831
0.564
c
SalI/SphI
5' BamHI
2.0 kb BamHI
5'
3'
EcoRI
HindIII
BamHI

4.2 GENOMIC ORGANISATION OF THE OVINE BLG GENE

Restriction enzyme digestions of sheep genomic DNA were carried out using some of the restriction enzymes which cut within the genomic clones. Southern blotting and hybridisation with a 424 bp fragment derived from the ovine BLG cDNA, p931 (see Gaye et al., 1986; figure 2, Ali and Clark, 1988), was used to show that the fragment sizes agree with sizes predicted from the restriction maps of the phage clones. This analysis would also show whether ovine BLG is encoded by a multi- or a single-copy gene. Figure 4.2a shows the band pattern seen for *HindIII*, *BamHI*, *EcoRI*, *PvuII* and *PstI* digests of sheep genomic DNA. The expected band pattern is seen in each case. For example, the *HindIII* digest yielded bands of 4.3 and 9.6 kb (fragments 2 and 3 respectively, in figure 4.2b, c), corresponding to the expected SS1- and SS12-like bands, respectively. This shows that the sheep DNA used here contains two different BLG gene types - one SS1-like and a SS12-like BLG gene. However, two bands of 6.7 kb and 7.8 kb are also seen. One (or two) bands, both greater than 7 kb in size, should be seen by the generation of a *HindIII* fragment on cutting at a *HindIII* site present 5' to the cloned SS1 and SS12 sequences. Thus, the 6.7 kb fragment is unexpected. Similarly, the *EcoRI* digestion shows a band (doublet) of 4.4 kb (fragments 6 in figure 4.2b, c) and an unexpected band of 7.4 kb. The other digests also yield bands whose presence is not predicted from the phage restriction maps. The expected bands are always seen and at an intensity consistent with a copy number of one (figure 4.2a - see digests probed with p931). Bell and McKenzie (1967) carried out segregation studies which show that ovine BLG is encoded by a single gene locus, with two alleles, A and B. My own data, presented in chapter 5, agrees with their findings.

Figure 4.2b shows the *HindIII*, *EcoRI* and *BamHI* fragments predicted from the phage maps (figure 4.2c). One (or two) other *HindIII* should also be present when this probe is used. In addition to the expected *EcoRI* and *BamHI* fragments (see above paragraph) and *PvuII* and *PstI* fragments, other bands are seen.

To further analyse the "related" sequences which hybridise to the cDNA, Southern blotted *HindIII*, *EcoRI* and *BamHI* digests of sheep genomic DNA were probed with subclones of SS1 (figure 4.2c). The *SphI/SalI* fragment contains 5' flanking sequences, the 5' *BamHI* fragment is 4.4 kb in length and contains about 1.5 kb of 5' flanking sequences and extends 3' into intron 3. The 2 kb *BamHI* fragment contains exons IV-VII and about 95 bp of 3' flanking sequence.

Hybridisation with the 2 kb *BamHI* fragment containing the BLG gene exons IV to VII (see section 4.3) is shown in figure 4.2a (and figure 4.2b). 4.3 kb and 9.6 kb (fragments 2 and 3, respectively) *HindIII* fragments, a single 4.4 kb *EcoRI* fragment (fragment 6) and a 2.0 kb *BamHI* fragment (fragment 9), were seen. Each digest also yielded other bands . Thus a 6.7 kb *HindIII* fragment, a 7.4 kb *EcoRI* fragment and 6.2, 6.6 and 9.2 kb *BamHI* fragments were seen. These fragments were also seen with the cDNA probe.

Two related functional gene sequences would be expected to show greater divergence (i.e. sequence dissimilarity) in non-transcribed (5' and 3' flanking sequences) and in intronic sequences, than in mRNA-encoding sequences. Therefore, sheep genomic DNA was probed with the 3.9 kb SS1 *SalI/SphI* fragment. This fragment contains no BLG gene transcribed sequences. As well as the fragment of expected size (fragment 1 - 7.8 kb) the *HindIII* digest contains a 3.0 kb fragment. It is more difficult to know which of the *EcoRI* and *BamHI* bands may not form part of the SS1- and SS12-like gene due to restriction sites present 5' of the cloning sites. The 4.4 kb *BamHI* band (fragment 8) is predicted from the phage map. The slightly

smaller 4.2 kb band probably arises from the presence of a *BamHI* site in sequences 5' of the SS1 and SS12 cloning sites (fragment 7). A 1.8 kb fragment is also present. The similar relative band intensities of the 4.2 and 4.4 kb fragments suggest that a *BamHI* site is present at the same position in SS1- and SS12-like genes 5' of the cloned sequences, indicating that the 1.8 kb fragment is not part of the BLG gene locus.

Probing with the 4.4 kb *BamHI* fragment, which contains exons I-III (including the exonic sequences encoding the amino-acids which are most highly conserved in the superfamily which includes BLG - see chapter 6), however, showed little evidence of "related" sequences. *HindIII* and *EcoRI* digests yielded only the expected fragments. The *BamHI* digestion, gave an extra fragment, which may be the same fragment noted with the *SphI/SalI* probe.

Sheep genomic DNA restriction digests probed with the BLG cDNA and with subclones of SS1 indicate that no major rearrangements have occurred in the cloning process. The *HindIII* digest results are consistent with SS1 and SS12 being alleles; 4.4 kb and 9.6 kb fragments were obtained when the cDNA or the 2 kb *BamHI* fragment (which is contained within the 4.4 and 9.6 kb *HindIII* fragments) were used as probes (also see chapter 5). Nevertheless, these results raise the possibility of an additional BLG gene. The "extra" bands obtained are of lower intensity than the expected ones. The simple band pattern (for example a single 6.7 kb *HindIII* fragment - figure 4.2a, 4.2b(i)) suggests the presence of only one other BLG-like gene. It seems improbable that another functional BLG gene is present since there is no evidence for its protein product in sheep milk. Results presented in chapter 3 demonstrate that the BLG gene is tissue-specific. Expression was not detected in any other tissue tested. Of course, not all tissues have been checked and the possibility of a BLG gene being expressed in another tissue cannot be ruled out. A human

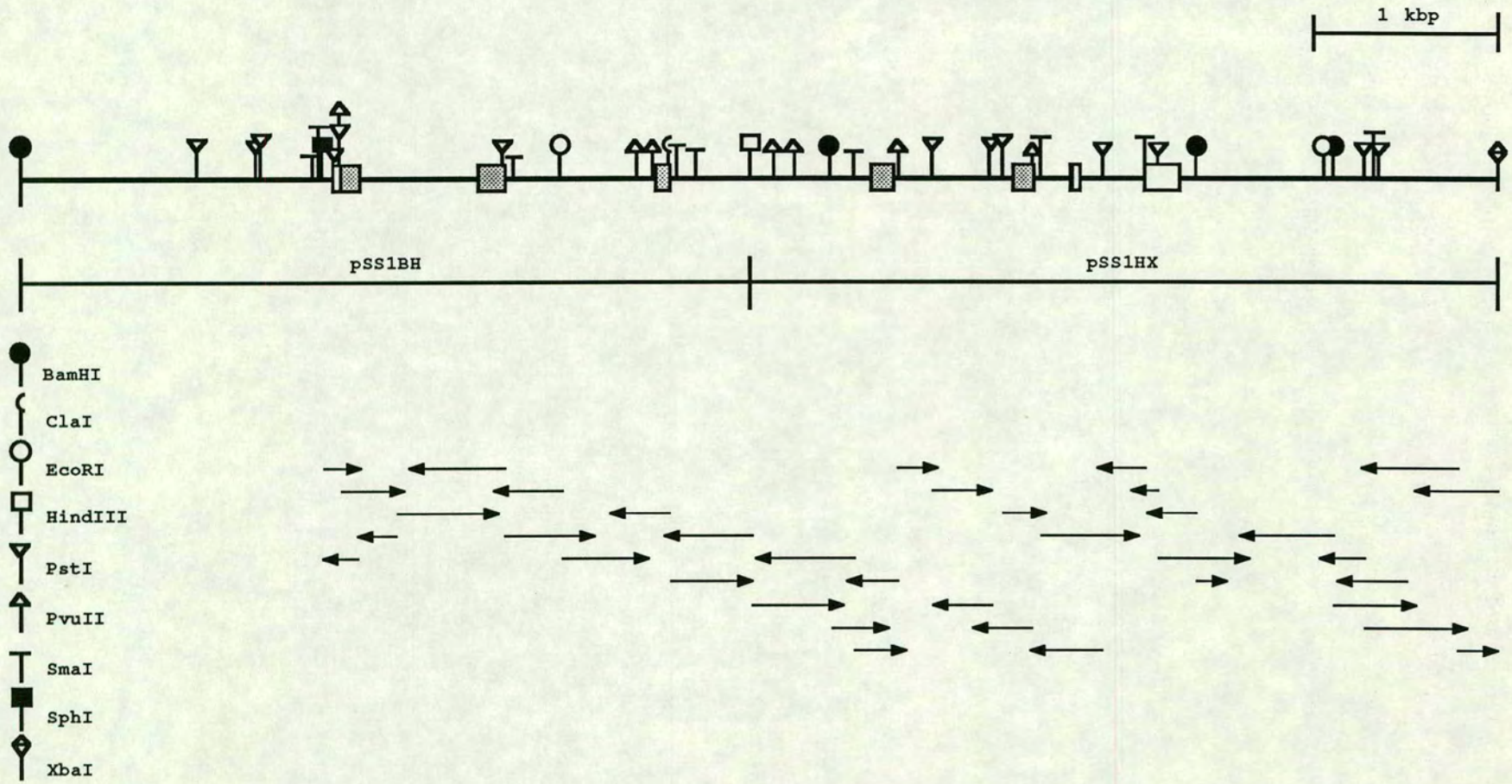
endometrial protein, α_2-pregnancy endometrial protein recently described (Huhtala et al., 1987; Bell et al., 1987), has been partially sequenced and shares 50% amino-acid homology with BLG. It is possible that a similar gene exists in sheep. Uterine RNA prepared from non-pregnant and early pregnancy ewes did not reveal an mRNA for such a gene (data not shown; see chapter 6). This, however, remains a possibility. Alternatively, these sequences may form a BLG pseudogene. The possibility of just one other BLG-like gene is confused by the observation of three large *BamHI* fragments with the 2 kb *BamHI* probe. More work is required to construct a restriction map of the "second" gene(s).

The sheep genomic DNA used in these digests was not the DNA used to make the genomic library. Identical results were obtained when the latter DNA was probed with p931 (A. J. Clark, unpublished results).

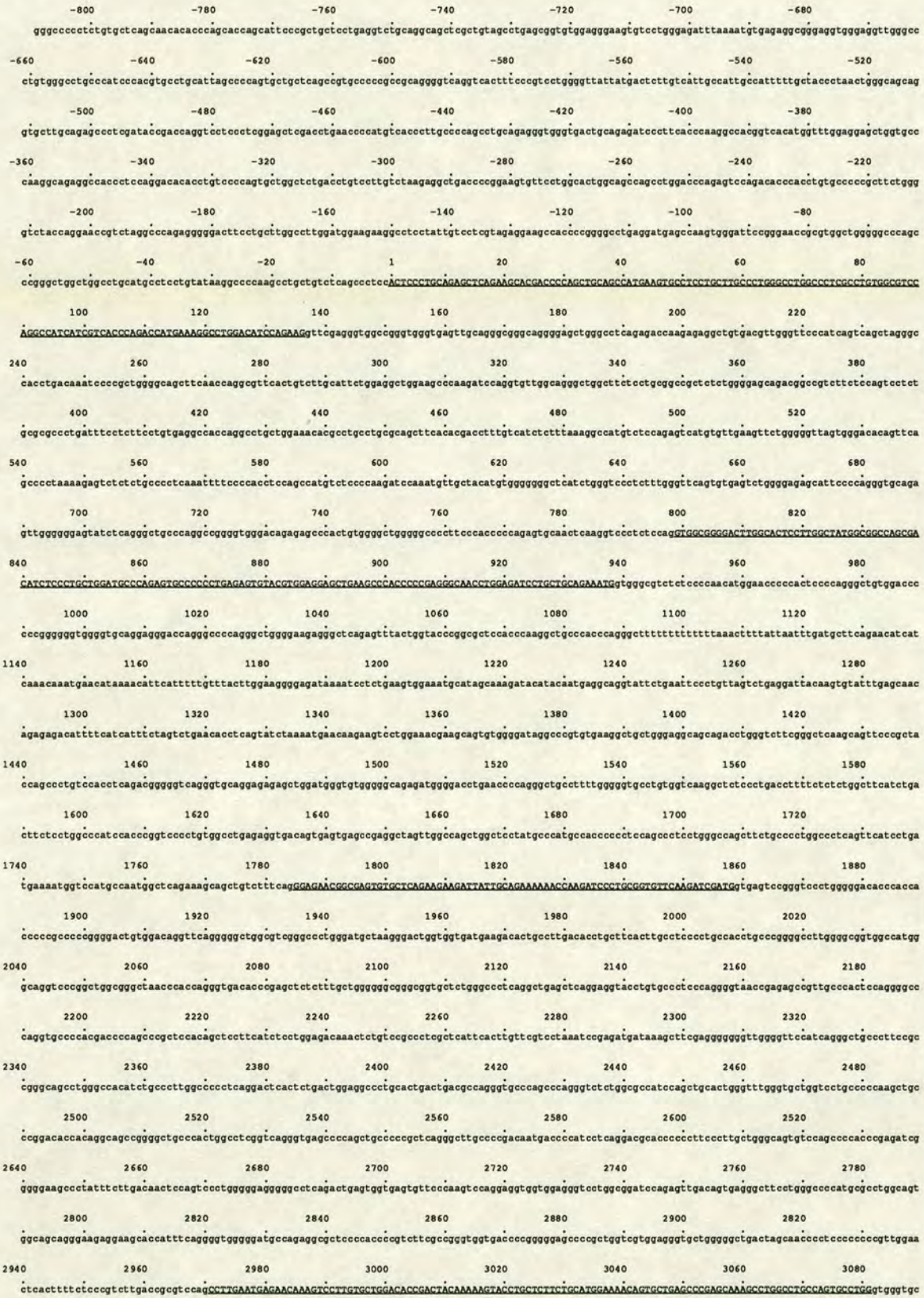
4.3 DNA SEQUENCING OF THE OVINE BLG GENE

The DNA sequence of SS1 has been determined from the *SphI* site 45 bp upstream of the transcriptional start site to the *XbaI* site about 1.9 kb downstream of the polyadenylation signal. A number of sub-clones were constructed and restriction mapped. These were used to derive restriction fragments for DNA sequencing. The DNA sequencing protocols used are described in the Materials and Methods section. Two subclones, pSS1BH and pSS1HX, were constructed in the plasmid vector ptg1(Poly) (Lathe et al., 1987). These were restriction mapped using a variety of restriction enzymes, including relatively rare-cutting enzymes whose presence within exons was predicted from the cDNA sequence of p931 (Gaye et al., 1986).

Figure 4.3 Structure of the BLG gene SS1. The figure shows the detailed restriction map of the SS1 subclones pSS1BH and pSS1HX. The restriction sites used are shown. The exons are shown as open boxes; coding regions are shaded. M13 clones are shown as arrows, denoting the extent and direction of sequence.


1 kbp
pSS1BH
pSS1HX
BamHI
ClaI
EcoRI
HindIII
PstI
PvuII
SmaI
SphI
XbaI

Ovine β-lactoglobulin Gene DNA Sequence



3100 3120 3140 3160 3180 3200 3220

caaccctggctgcccagggagaccagctgcgtggtccttgctgcaacaggggtgggggtgggagcttgatccccaggaggaggaggggtgggggtccctgagtcccgccaggagagagtggtcgcataccgggagccagtctgctgt

3240 3260 3280 3300 3320 3340 3360 3480

gggcctgtgggtggctgggacgggggccagacacacaggccgggagacgggtgggctgcagaactgtgactggtgtgaccgtcgcgatgggccggtggtcactgaatctaacagcctttgttaccgggggagtttcaattatttcccaa

3400 3420 3440 3460 3480 3500 3520

aataagaactcaggtacaaagccatctttcaactatcacatcctgaaaacaaatggcaggtgacattttctgtgccgtagcagtcccactgggcattttcagggcccctgtgccagggggggcgcgggcatcggcgagtggaggctcctgg

3540 3560 3580 3600 3620 3640 3660 3780

ctgtgtcagccggcccaggggagaggaagggacccggacagccagaggtgggggcaggctttcccccctgtgacctgcagacccactgcactgccctgggaggaagggaggggaactaggccaaggggggaagggcaggtgctctggagggc

3700 3720 3740 3760 3780 3800 3820

aagggcagacctgcagaccacccctgggagcagggactgacccccgtccctgccccatagTCAGGACCCCGGAGGTGGACAACGAGGCCCTGGAGAAATTCGACAAAGCCCTCAAGGCCCTGCCCATGCACATCCGGCTTGCCTTCAACC

3840 3860 3880 3900 3920 3940 3960 3980

CGACCCAGCTGGAGGGtgagcacccaggccccgcccttccccagggcaggagccacccggcccgggacgacctcctcccatggtgacccccagctccccaggcctcccaggaggaaggggtggggtgcagcaccccgtgggggccccct

4000 4020 4040 4060 4080 4100 4120

ccccacccccctgccaggcctctcttcccgaggtgtccagtccatcctgacccccccatgactctccctccccacagGGCAGTGCCACGTCTAGGTGAGCCCCTGCCGGTGCCTCTGGGGtaagctgcctgccctgccccacgtcctgg

4140 4160 4180 4200 4220 4240 4260 4280

gcacacacatggggtaggggtcttggtggggcctgggaccccacatcaggccctggggtccccccctgtgagaatggctggaagctggggtccctcctggcgactgcagagctggctggccgcgtgccactcttgtgggtgacctgtgtc

4300 4320 4340 4360 4380 4400 4420

ctggcctcacacactgacctcctccagctccttccagcagagctaaggctaagtgagccagaatggtacctaaggggaggctagcggtccttctcccgaggaggggctgtcctggaaccaccagccatggagaggctggcaagggtctgg

4440 4460 4480 4500 4520 4540 4560 4580

caggtgccccaggaatcacaggggggccccatgtccatttcagGGGCCCGGGAGCCTTGGACTCCTCTGGGGACAGACGACGTCACCACCGCCCCCCCCCCATCAGGGGGACTAGAAGGGACCAGGACTGCAGTCACCCTTCCTGGGACCC

4600 4620 4640 4660 4680 4700 4720

AGGCCCCTCCAGGCCCCTCCIGGGGCTCCTGCTCTGGGCAGCTTCTCCTTCACCAATAAAGGCATAAACCTGTgctctcccttctgagtctttgctggacgacgggcaggggggtggagaagtggtggggagggagtctggctcagaggat

4740 4760 4780 4800 4820 4840 4860 4880

gacagcggggctgggatccagggcgtctgcatcacagtcttgtgacaactgggggcccacacacatcactgcggctctttgaaactttcaggaaccagggagggactcggcagagacatctgccagttcacttggagtgttcagtcaaca

4900 4920 4940 4960 4980 5000 5020

cccaaactcgacaaaggacagaaagtggaaaatggctgtctcttagtctaataaatattgatatgaaactcaagttgctcatggatcaatatgcctttatgatccagccagccactactgtcgtatcaactcatgtacccaaacgcactg

5040 5060 5080 5100 5120 5140 5160 5180

atctgtctggctaatgatgagagattcccagtagagagctggcaagaggtcacagtgagaactgtctgcacacacagcagagtccaccagtcatcctaaggagatcagtcctggtgttcattggaggactgatgttgaagctgaaactcc

5200 5220 5240 5260 5280 5300 5320

aatgctttggccacctgatgtgaagagctgactcatttgaaaagaccctgatgctgggaaagattgagggcaggaggagaagggacgacagaggatgagatggttggatggcatcaccaacacaatggacatgggtttgggtggactcc

5340 5360 5380 5400 5420 5440 5460 5480

aggagttggtgatggacagggaggcctggcgtgctacggaagcggtttatggggtcacaaagactgagtgactgaactgagctgaactgaatggaaatgaggtatacagcaaagtggggatttttagataataagaatatacacataac

5500 5520 5540 5560 5580 5600 5620

atagtgtatactcatatttttatgcatacctgaatgctcagtcactcagtcgtatctgactctgtgacctatggaccgtagccttccaggtttcttctgtccacagaatttctccaaggcaagaatactggagtgggtagccatttcctcc

5640 5660 5680 5600 5720 5740 5760 5780

tccaggggatcctcccgacccagggattgaaccggcatctcctgtattggcaggtggattctttaccactgtgccaccagggaagcccgtgttactctctatgtcccacttaattaccaaagctgctccaagaaaaagccctgtgccct

5800 5820 5840 5860 5880 5900 5920

ctgagcttcccggcctgcagagggtggtgggggtagactgtgacctgggaacaccctcccgcttcaggactcccgggccacgtgacccacagtcctgcagacagccgggtagctctgctcttcaaggctcattatctttaaaaaaaactg

5940 5960 5980 6000 6020 6040 6060 6080

aggtctatttgtgacttcgctgccgtaacttctgaacatccagtgcgatggacaggacctcctccccaggcctcaggggcttcagggagccagccttcacctatgagtcaccagacactcgggggtggcccgccttcagggtgctcac

6100 6120 6140 6160 6180 6200 6220

agtcttcccatcgtcctgatcaaagagcaagaccaatgacttcttaggagcaagcagacacccacaggacactgaggttcaccagagctgagctgtcttttgaacctaaagacacacagctctcgaaggttttctctttaatctggatt

6240 6260 6280 6300 6320 6340 6360 6380

taaggcctacttgcccctcaagagggaagacagtcctgcatgtccccaggacagccactcggtggcatccgaggccacttagtattatctgaccgcacccctggaattaatcggtccaaactggacaaaaaccttggtgggaagtttcatc

6400 6420 6440 6460 6480 6500 6520

ccagaggcctcaaccatcctgctttgaccaccctgcatctttttttcttttatgtgtatgcatgtatatatatatatatttttttttttttcatttttggctgtgctggctgttcgttgcagttcggtgcgcaggcttctctctagt

6540 6560

ttctctctagtcttctcttatcacagagcagtctctaga

Figure 4.4 DNA sequence of the BLG gene SS1. The sequence presented includes the 5' flanking sequences (-810 to -40), determined by S. Anderson (Harris et al., 1988). The sequence of the coding strand is shown. Exon sequences are shown in upper case and are underlined. 5' and 3' flanking and intronic sequences are in lower case. The numbers refer to distance from the transcriptional start site.

Figure 4.5 Exonic sequence of the BLG gene SS1. Exon sequences are shown in upper case, flanking sequences are in lower case. The predicted amino-acid sequence is shown above the DNA sequence. The numbers immediately above the amino-acid sequence are given relative to the N-terminal amino-acid of the mature BLG polypeptide. Negative numbers refer to the BLG gene signal peptide. The numbers immediately below the DNA sequence are distances from the transcriptional start site (see also figure 4.4). Differences from the cDNA sequence (Gaye et al., 1986) are shown immediately below the DNA sequence. (*) below the DNA sequence indicates a gap in the alignment with the cDNA sequence. Putative TATA, transcriptional start site, translation start and stop and AATAAA signals are underlined (see table 4.1). The single amino-acid difference between BLG-A (Kolde and Braunitzer, 1983a) and SS1 sequence is shown (amino-acid 20).

Exon I 136 bp

```
gcatgcctcctgtataaggccccaagcctgctgtctcagccctccACTCCCTGCAGAGCTCAGAAGCACGACCCCAGCTG
                                                    1          T
     -18                                                      -1 +1
     MetLysCysLeuLeuLeuAlaLeuGlyLeuAlaLeuAlaCysGlyValGlnAlaIleIleValThrGlnThrMet
CAGCCATGAAGTGCCTCCTGCTTGCCCTGGGCCTGGCCCTCGCCTGTGGCGTCCAGGCCATCATCGTCACCCAGACCATG
                                                                          100
      10
LysGlyLeuAspIleGlnLys
AAAGGCCTGGACATCCAGAAGgttcgagggt
```

Exon II 140 bp

```
                          20                                  30
          ValAlaGlyThrTrpHisSerLeuAlaMetAlaAlaSerAspIleSerLeuLeuAspAlaGlnSerAlaP
ccctctccagGTGGCGGGGACTTGGCACTCCTTGGCTATGGCGGCCAGCGACATCTCCCTGCTGGATGCCCAGAGTGCCC
        800             Tyr
     40                                50                                  60
roLeuArgValTyrVAlGluGluLeuLysProThrProGluGlyAsnLeuGluIleLeuLeuGlnLysTr
CCCTGAGAGTGTACGTGGAGGAGCTGAAGCCCACCCCCGAGGGCAACCTGGAGATCCTGCTGCAGAAATGgtgggcgtct
                               900
```

Exon III 74 bp

```
                                   70                                  80
          pGluAsnGlyGluCysAlaGlnLysLysIleIleAlaGluLysThrLysIleProAlaValPheLysIle
tgtctttcagGGAGAACGGCGAGTGTGCTCAGAAGAAGATTATTGCAGAAAAAACCAAGATCCCTGCGGTGTTCAAGATC
                         1800

AspA
GATGgtgagtccgg
Asn
```

Exon IV 111 bp

```
                     90                                  100
          laLeuAsnGluAsnLysValLeuValLeuAspThrAspTyrLysLysTyrLeuLeuPheCysMetGluAs
ccgcgtccagCCTTGAATGAGAACAAAGTCCTTGTGCTGGACACCGACTACAAAAAGTACCTGCTCTTCTGCATGGAAAA
                                   3000
 110                          120
nSerAlaGluProGluGlnSerLeuAlaCysGlnCysLeuV
CAGTGCTGAGCCCGAGCAAAGCCTGGCCTGCCAGTGCCTGGgtgggtgcca
```

Exon V 105 bp

```
                                130                                 140
          alArgThrProGluValAspAsnGluAlaLeuGluLysPheAspLysAlaLeuLysAlaLeuProMetHi
tgccccatagTCAGGACCCCGGAGGTGGACAACGAGGCCCTGGAGAAATTCGACAAAGCCCTCAAGGCCCTGCCCATGCA
                                                          3800
          150
sIleArgLeuAlaPheAsnProThrGlnLeuGluG
CATCCGGCTTGCCTTCAACCCGACCCAGCTGGAGGgtgacgaccc
```

Exon VI 42 bp

```
               160
          lyGlnCysHisVal***
tcccccacagGGCAGTGCCACGTCTAGGTGAGCCCCTGCCGGTGCCTCTGGGgtaagctgct
                                    4100
```

Exon VII 180 bp

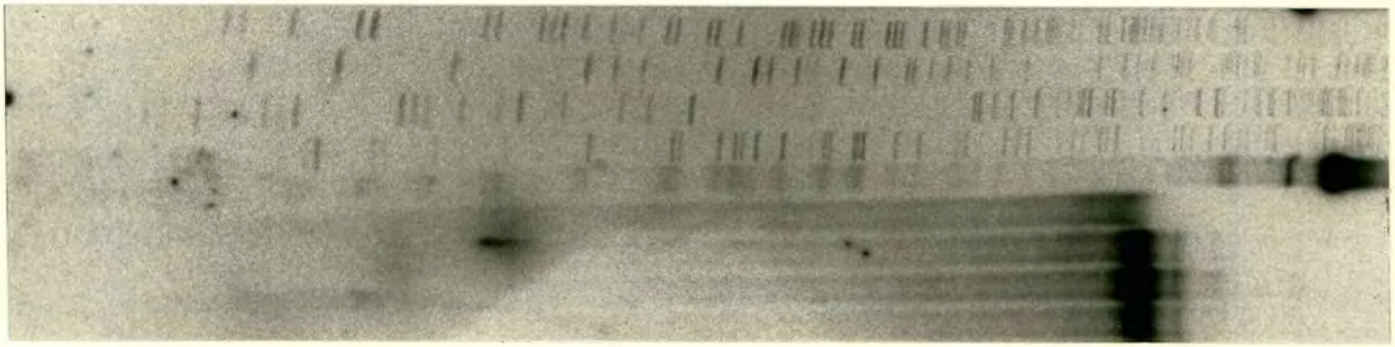
```
ccattttcagGGCCCGGGAGCCTTGGCTCCTCTGGGGACAGACGACGTCACCACCGCCCCCCCCCCATCAGGGGGACTA
                          4500                              *

GAAGGGACCAGGACTGCAGTCACCCTTCCTGGGACCCAGGCCCCTCCAGGCCCCTCCTGGGGCTCCTGCTCTGGGCAGCT
                                          4600

TCTCCTTCACCAATAAAGGCATAAACCTGTgctctcccttctgagtctttgctggacgacgggcaggggggt
                                                                  4700
```

Table 4.1. *DNA sequence signals present in SS1*

Signal	*Gene*	*Sequence*	*Position*
Transcription initiation	BLG	cTGTATAAgGCc	-33
	Consensus[a]	GNGTATAWAWNG	-30
	BLG	CACTCC	+1
	Consensus[b]	CANYYY	+1
Translation initiation	BLG	CaGCCATG	+41
	Consensus[c]	CCRCCATG	
Translation termination	BLG	TAG	+4082
Donor/lariat/acceptor splice sites			
Donor splice sites	Intron 1	GTtcGa	+136
	Intron 2	GTGgGc	+938
	Intron 3	GTGAGT	+1858
	Intron 4	GTGgGT	+3081
	Intron 5	GTGAGc	+3854
	Intron 6	GTAAGc	+4109
	Consensus[a]	GTRAGT	
Acceptor splice sites	Intron 1	CCCTCTCCAG	+799
	Intron 2	TgTCTTTCAG	+1784
	Intron 3	CCgCgTCCAG	+2970
	Intron 4	TgCCCCAtAG	+3749
	Intron 5	TCCCCCACAG	+4067
	Intron 6	TCCaTTTCAG	+4482
	Consensus[a]	YYYYYYNCAG	
"Lariat" splice sites	Intron 1	CccAC	+766
	Intron 2	CTGAT	+1739
	Intron 3	CTcAC	+2943
	Intron 4	CTGAC	+3729
	Intron 5	CTGAC	+4039
	Intron 6	CacAg	+4459
	Consensus[d]	CTRAY	
Termination signals	BLG	AATAAA	+4644
	Consensus[e]	AATAAA	
	BLG	TGaGTcTT	+4674
	Consensus[f]	YGTGTTYY	

Underlined bases refer to position relative to the transcriptional start site. The BLG sequences are shown in upper case where there is agreement with the consensus and in lower case when it differs from the consensus. R is A or G, Y is C or T.
Consensus signals were taken from [a]Breathnach and Chambon (1981), [b]Bucher and Trifonov (1986), [c]Kozak (1984), [d]Keller and Noon (1984), [e]Proudfoot and Brownlee (1976), [f]McLauchlan et al. (1985).

Figure 4.6 S1 protection. A labelled fragment containing SS1 sequences from a *SphI* site (-41) to a *TaqI* site (+141) was used as probe. The figure shows the sequence ladder (G, A, T, C) of this fragment cloned into M13. The track labelled probe, contains the ^{32}P-labelled *SphI/TaqI* fragment (+ M13 polylinker 5' to the *SphI* site - see Materials and Methods). The next lane shows S1 protection of tRNA (S1 protection control). Lanes 1, 2 and 3 contain ovine lactating mammary gland RNA. S1 protection of mammary RNA yields bands 136 nucleotides and 145 nucleotides long, which position the major cap site, as shown in figure 4.5.


G
A
T
C
Probe
tRNA
1
2 Sheep mam. RNA
3
136 nucl.

Cloning DNA fragments generated using these restriction enzymes into the polylinker of bacteriophage M13 vectors, tg130 and tg131 (Kieny et al., 1983), allowed sequencing of exonic regions and the exact determination of exon/intron junctions. The sequence and restriction digest information was used to determine which other fragments to sequence in order to obtain the complete BLG gene DNA sequence (also see figure 4.2; Ali and Clark, 1988).

Figure 4.3 shows a restriction map of pSS1BH and pSS1HX together with the determined exonic structure. Also shown are the M13 clones which were used for sequencing, the extent and direction of sequence obtained is shown by the length and direction of the arrows. Greater than 85% of the sequence has been determined on both strands. Figure 4.4 shows the entire ovine BLG gene (SS1) sequence which includes 760 bp of 5' flanking DNA sequence starting just downstream of an *AvaI* site, to the *SphI* site (this sequence was obtained by S. Anderson, Hons. project (1988) - Dept. of Genetics, Edinburgh University). This 5' sequence is complete on both strands (Harris et al., 1988).

The ovine BLG gene transcription unit is 4.9 kb long, with seven exons. Translation starts in exon I and ends in exon VI. Exon VII is non-coding. The translated sequences are identical with those obtained by Gaye et al. (1986) for the mRNA. SS1 encodes a gene for ovine BLG-B (a histidine residue is present at amino-acid 20) rather than the A-variant (a tyrosine residue at amino-acid 20) (Kolde and Braunitzer, 1983b; Gaye et al., 1986). Two differences between the cDNA sequence and the SS1 sequence, were noted. One is the presence of a C, instead of a T, in SS1 at position +14 in the 5' untranslated region. The second is an extra C in SS1 at +4530 (exon VII). This difference is found in a run of 10 Cs and may have been caused by a cloning error in SS1 or a reverse transcriptase error in the cDNA (although two independent cDNAs were sequenced (Gaye et al., 1986)). Figure 4.5

shows the sequence of the BLG gene exons, as well as splice junctions.

The transcriptional start site was tentatively placed 40 bp upstream of the A of the initiation codon, ATG. The putative transcription start site (...ccctccACTCCCT...) agrees with the consensus derived by Breathnach and Chambon (1981) (see Table 4.1). This putative start site is present 33 bp downstream of the only good TATA element-like sequence present in the region (see Table 4.1). Gaye et al. (1986) used primer extension to extend the mRNA sequence to the C at +2, consistent with the SS1 DNA sequence.

S1 mapping was carried out using total RNA prepared from a lactating sheep mammary gland (Material and Methods). The probe used to hybridise against mammary gland RNA extends from a *TaqI* site 4 bp downstream of the exon I/intron I splice junction, to the *SphI* site (-45). The main band obtained was 136 nucleotides long and shows that the major transcriptional start site is indeed as predicted (figure 4.6 and table 4.1). A minor start site was found to be present 9 bp upstream of this site. The sequence around this site (...tgtctcAGCCCTC...), also agrees with the consensus sequence derived by Breathnach and Chambon (1981). Similarly, RNase protection shows that minor transcription start sites are used. Nevertheless, the major transcription start site is the one seen in the S1 analysis (P. Brown, unpublished results). In figure 4.6, the major and minor transcriptional starts site are labelled. A "smear" was seen in the control (tRNA) track. It is not clear why this was seen, but it may indicate the presence of secondary structure in the probe, which is only slowly digested by S1. The 136 nucleotide fragment in the mammary RNA samples should also be digested in a similar manner to the control sample if it does not represent a true RNA-DNA hybrid. Furthermore, although the "probe" sample was incubated under identical conditions to the other samples, except that no S1 was added, the 136 nucleotide fragment should be seen if it was due to secondary

structure formation. The probe isolation method also meant that it was single-stranded, so DNA-DNA hybridisation could not occur. Thus, the available evidence suggests that this fragment does indeed show the transcriptional initiation sites of the ovine BLG gene.

Sequences at exon/intron junctions are usually well conserved and consensus sequences have been derived for splice donor and acceptor sites (Breathnach and Chambon, 1981). It is now known that an intermediate in splicing involves the formation of a "lariat", involving sequences 20-55 nucleotides upstream of the splice acceptor site (for review see Padgett et al., 1986). The sequences show some specificity (Keller and Noon, 1984). Sequences around the donor and acceptor splice sites of the six introns of SS1 agree well with the above consensus sequences. Sequences at the splice junctions of the ovine BLG gene conform well to Breathnach and Chambon's (1981) derived consensus sequences (table 4.1). The appropriate region of each intron was searched for putative Keller and Noon (1984) splice "lariat" sequences. Three of the six introns contain a sequence CTGAC, in absolute agreement with the consensus sequence. The other three introns have sequences differing from the consensus at one or more positions. These are shown in table 4.1.

Table 4.1 also shows the presumed translation start and stop codons. The positions of these codons correlate exactly with the amino-acid sequences of the mature BLG polypeptide sequenced by Kolde and Braunitzer (1983a) and the signal peptide sequence determined by Mercier et al. (1978). Thus the ovine BLG gene encodes a 180 amino-acid polypeptide which includes an 18 amino-acid signal peptide and a 162 amino-acid mature polypeptide.

Most messenger RNAs of RNA polymerase II-transcribed genes have a poly(A) tail. A highly conserved signal (AATAAA) is required for polyadenylation

(Proudfoot and Brownlee, 1976; Fitzgerald and Shenk, 1981). This sequence is most often present 20-30 bp upstream of the polyadenylation site. A point mutation in this sequence leads to a loss of correctly terminated mRNAs (Montell et al., 1983). It is not the only sequence required for termination as it is sometimes also found elsewhere within mRNAs. Further work has shown that other sequences near the AATAAA sequence are important for transcription termination. In particular, a sequence similar to YGTGTTYY is present 24-38 bp downstream from the AATAAA signal in 67% of mammalian and eukaryotic viral genes examined (McLauchlan et al., 1985). Deletion of this sequence from a HSV 'terminator' fragment linked downstream of the bacterial chloramphenicol transferase (CAT) gene gave significantly reduced levels of CAT activities and CAT mRNA 3' termini (McLauchlan et al., 1985).

As shown in table 4.1 and figure 4.4, the sequence AATAAA is present at position +4644, 19 bp 5' of the polyadenylation site. A sequence TGAGTCTT (at +4674) is similar to the YGTGTTYY motif. This motif is found within a stretch of about 20 Cs and Ts running 3' from the polyadenylation site. This kind of CT stretch is commonly found in sequences 3' of the polyadenylation site (Nussinov, 1986).

The BLG gene encoded within SS1 contains the sequence motifs which are present in most functional RNA polymerase II genes, as tabulated in table 4.1. In a following section I have attempted to search the gene sequences for the presence of sequence motifs which may regulate transcription of the ovine BLG gene.

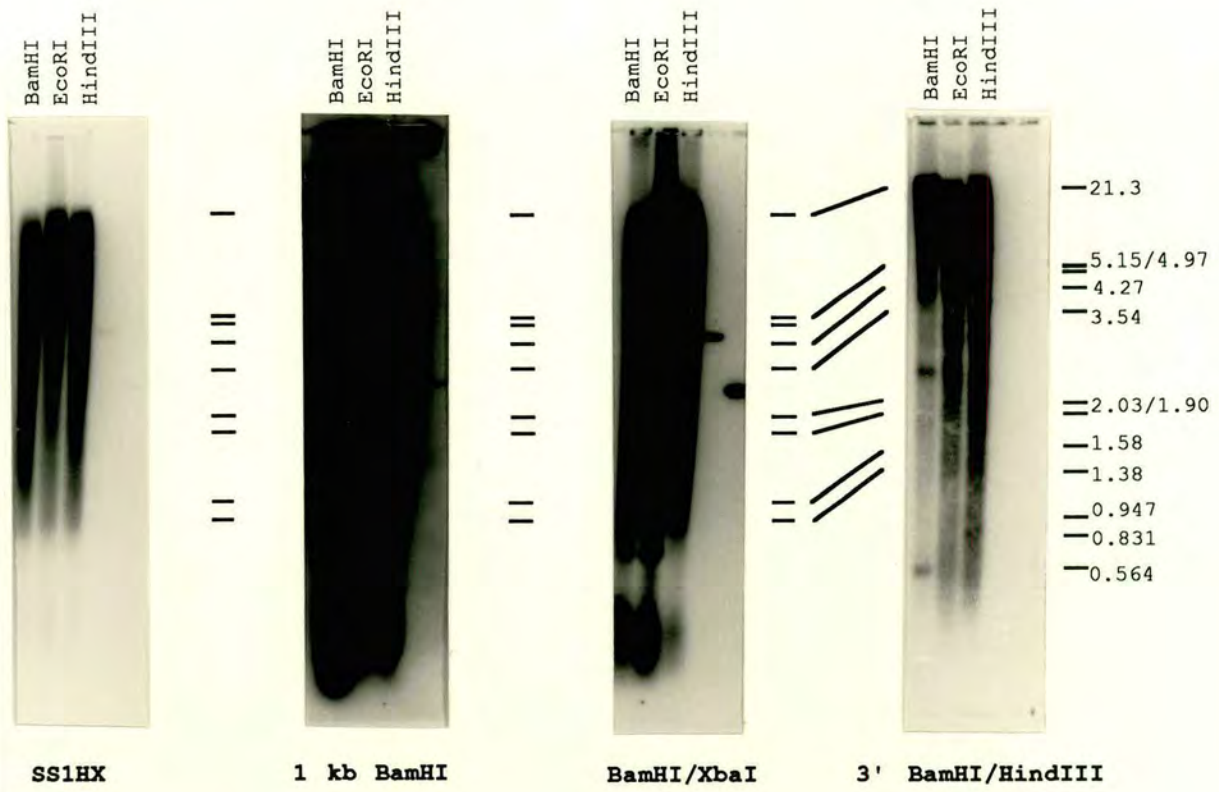
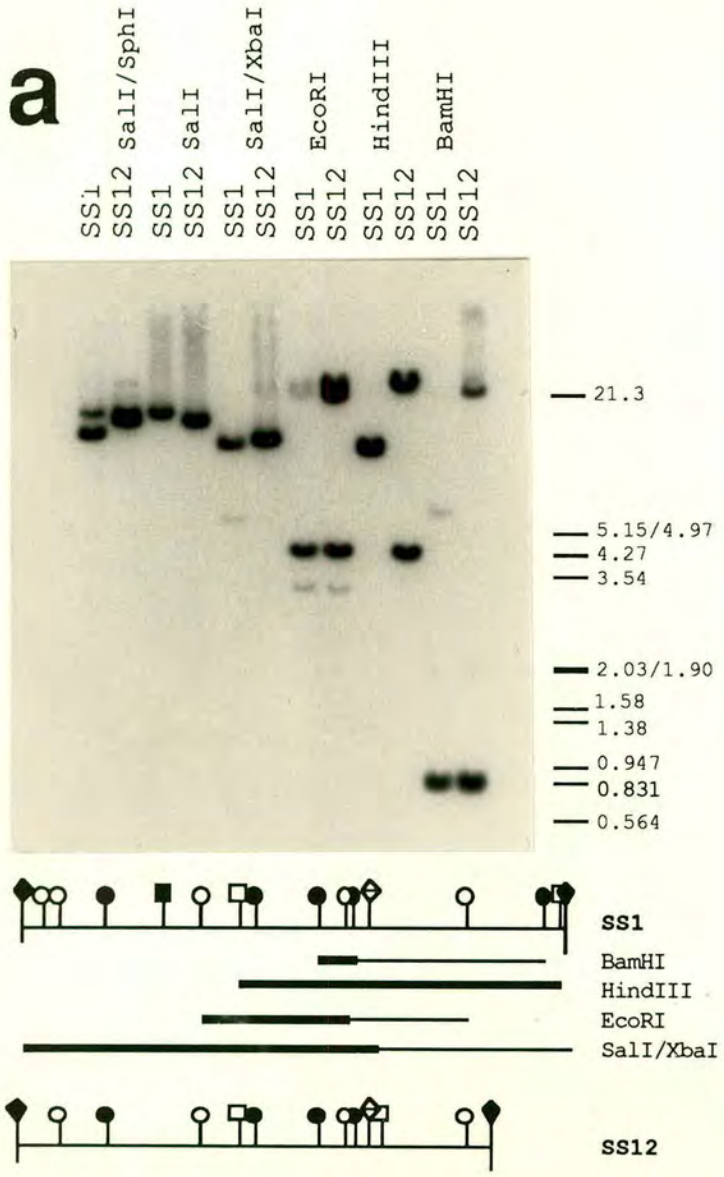
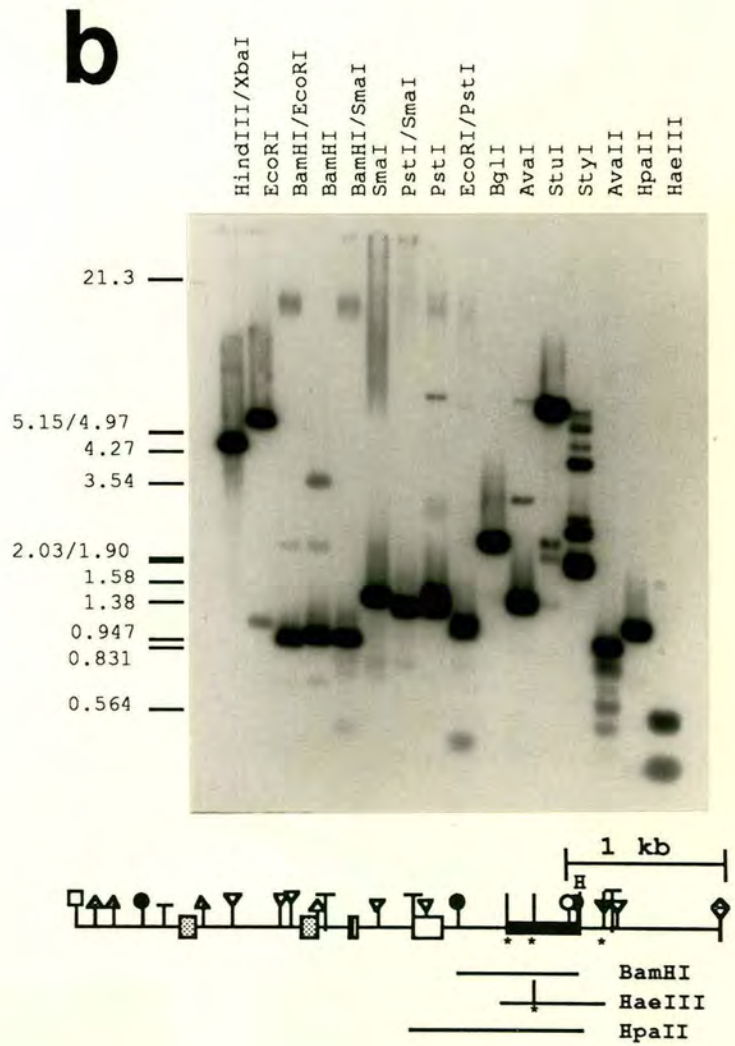

BamHI
EcoRI
HindIII
BamHI
EcoRI
HindIII
BamHI
EcoRI
HindIII
BamHI
EcoRI
HindIII
21.3
5.15/4.97
4.27
3.54
2.03/1.90
1.58
1.38
0.947
0.831
0.564
SS1HX
1 kb BamHI
BamHI/XbaI
3' BamHI/HindIII

Figure 4.7 BLG gene SS1 repeats. pSS1HX and derived fragments (~1 kb *BamHI* fragment just 3' of exon VII and *BamHI/XbaI* (+5650 to +6577)) were used to probe *BamHI*, *EcoRI* and *HindIII* digests of ovine genomic DNA. The 0.8 kb *BamHI/HindIII* fragment at the 3' end of SS1 was also used (3' *BamHI/HindII* - see figure 4.8). The numbers are lambda marker sizes, in kb.

Figure 4.8 Mapping BLG gene repeats. SS1 and SS12 (a) and pSS1HX (b) restriction digests were probed with ^{32}P-labelled ovine genomic DNA. The hybridising bands have been mapped below. In the map in (a), the heavy lines indicate strong hybridisation. In (b), only the most informative digests are shown schematically. The numbers are lambda marker sizes, in kb.


a
SS1
SS12
SalI/SphI
SS1
SS12
SalI
SS1
SS12
SalI/XbaI
SS1
SS12
EcoRI
SS1
SS12
HindIII
SS1
SS12
BamHI
21.3
5.15/4.97
4.27
3.54
2.03/1.90
1.58
1.38
0.947
0.831
0.564
SS1
BamHI
HindIII
EcoRI
SalI/XbaI
SS12


b
HindIII/XbaI
EcoRI
BamHI/EcoRI
BamHI
BamHI/SmaI
SmaI
PstI/SmaI
PstI
EcoRI/PstI
BglI
AvaI
StuI
StyI
AvaII
HpaII
HaeIII
21.3
5.15/4.97
4.27
3.54
2.03/1.90
1.58
1.38
0.947
0.831
0.564
1 kb
H
BamHI
HaeIII
HpaII

4.4 THE OVINE BLG GENE CONTAINS REPEATS

As well as unique sequences (e.g. the BLG gene) eukaryotic genomes contain sequences which are repeated many times throughout the genome. Functions for such repeats are not clear although a number of possible functions have been attributed to them; such as being origins of DNA replication or promoters of transcription. Two types of repeats are observed - the so-called satellite DNAs which consist of very long arrays of short sequences repeated many times. These repeated sequences themselves are often composed of multiple copies of shorter sequences. Digestion of genomic DNA with a restriction enzyme which cuts once within a satellite repeat often gives sharp bands of unit length and may often give a ladder of fragments differing in size by unit lengths. The other type of repeat sequences are typified by the human Alu repeats. These are short sequences about 300 bp in length. A total of about 300,000 Alu repeats are scattered throughout the human genome. Similar repeats have been cloned from other mammals, including rodents and cattle. These repeats sometimes contain sequences capable of acting as promoters recognised by RNA polymerase III and many have been transcribed *in vitro* (for reviews see Lewin, 1980; Singer, 1981).

When restriction enzyme digests of sheep genomic DNA were probed with the 10.5 kb SS1 *SalI/XbaI* fragment a general smear diagnostic of the presence of repeat(s) was seen (A. J. Clark, unpublished observation). SS1 sub-clones were used to map a repeat(s) to the 2 kb region 5' of the *XbaI* site (figures 4.2 and 4.7). Other repeats were mapped to the extreme 3' end of SS1 (figure 4.7 - 3' *BamHI/HindIII* probe). In order to confirm these results restriction enzyme digests of SS1 and SS12 were probed with radioactively labelled sheep genomic DNA. 100 ng

of sheep genomic DNA restricted with *HindIII* was ^{32}P-labelled, using oligo-labelling (see Materials and Methods). Repeat sequences are present in many thousands of copies. For example, 100 ng of human genomic DNA would contain about 1 ng of Alu repeat sequences but only about 0.1 pg of a 5 kb single copy gene of the same size as the ovine BLG gene sequences. Probing with 100 ng of sheep genomic DNA should indicate restriction fragments which contain repeats. Restriction enzyme digests of the subclone, pSS1HX, were also probed with sheep genomic DNA as it had already been established that sequences within this subclone contain a repeat(s) (figure 4.7).

Figure 4.8a,b show the results of this experiment. SS1 and SS12 were digested with *BamHI*, *EcoRI*, *HindIII* and *SalI/XbaI*. The restriction fragments which hybridise to the sheep genomic DNA are also mapped in figure 4.8. This shows that a repeat is present in both SS1 and SS12 in the 2 kb region, with its 3' end at *XbaI*. These results indicate that the region 5' of +5808 (*PstI* site) contains all repeat sequences (figure 4.8b). Since the 900 bp *BamHI* fragment does not hybridise to the bands, sequences between +5650 (*BamHI* site) and +5804 probably contain much of the repeats.

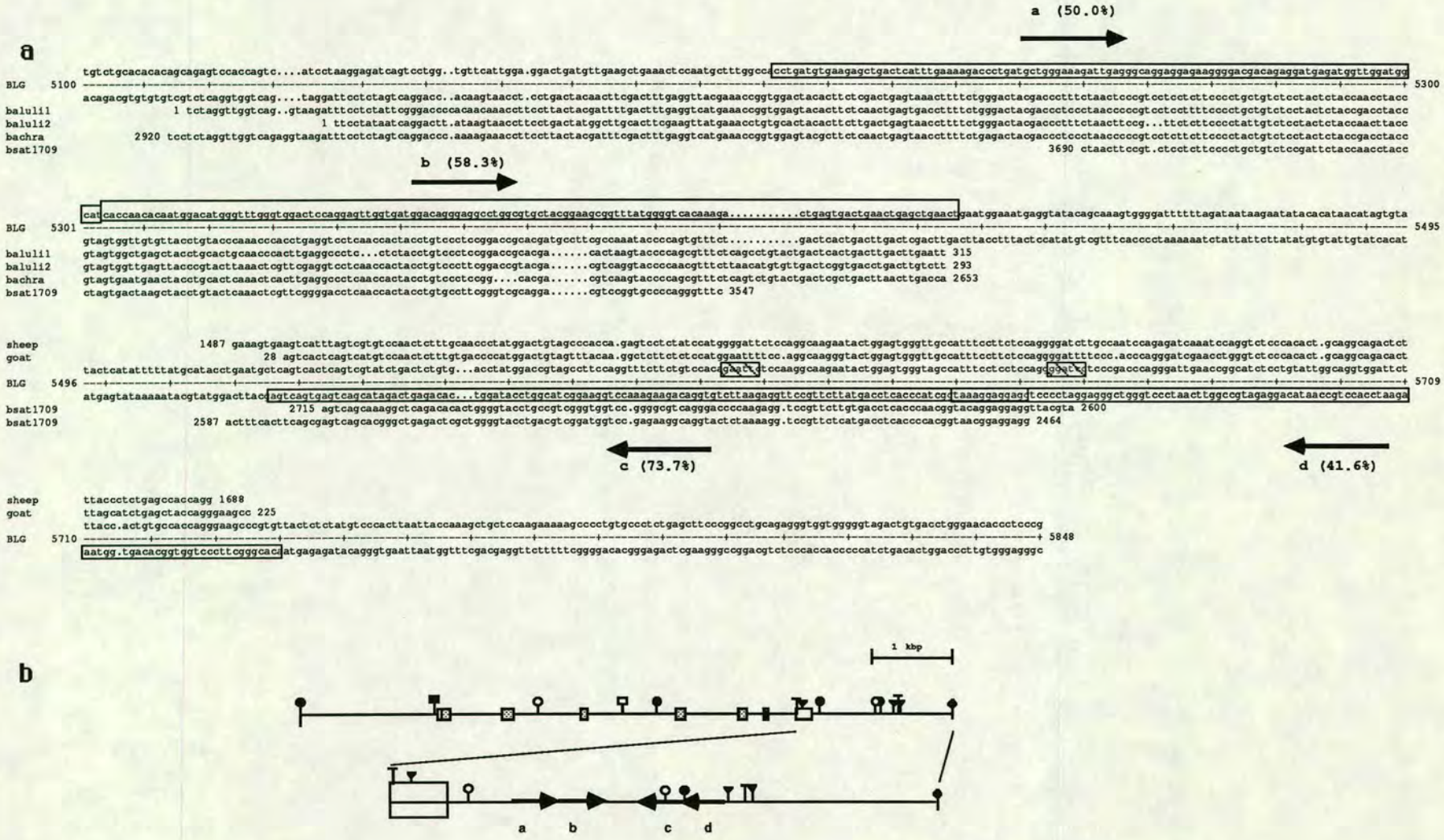
Other repeats also appear to be present, one in SS1, the other in SS12. The second repeat present in SS1 appears to reside at the 3' end of the phage clone, a region which is absent from SS12. Figure 4.7 shows that sheep genomic DNA probed with the 3' 0.8 kb *BamHI/HindIII* fragment gave a smear consistent with the presence of repeats. The second repeat present in SS12 appears to reside at the 5' end. SS12 barely extends further 5' than SS1, suggesting that this repeat may be absent from SS1. The repeat may have arisen by an insertional/deletional event in either SS1 or SS12. Possible evidence for this comes from an *EcoRI* site which is present at the 5' end of SS1 but absent from SS12. The presence of three repeats

within about 15 kb is not unusual (for example, see Watanabe et al., 1982).

4.5 SEQUENCE SIMILARITIES WITH OTHER REPEATS

The University of Wisconsin program WORDSEARCH (Devereux et al., 1984) was used to compare BLG gene sequences against both the GENBANK and EMBL DNA sequence databases. A number of bovine, sheep and goat repetitive sequences share sequence similarity with the 3' flanking sequences. An alignment of BLG gene sequences with these repeats is shown in figure 4.9a. Three of the repeat sequences are Alu-like bovine repeats, estimated to be present in at least 100,000 copies in the cow genome (Watanabe et al, 1982; Duncan, 1987). Watanabe et al. (1982) have shown that these repeats are distantly related to the human Alu repeats. They also showed that the repeat consists of 120 bp unit sequences in tandem. Skowronski et al. (1984) have derived a consensus sequence for the basic unit. Comparison of the BLG repeats with the consensus sequence shows four copies of the 120 bp repeat unit are present. These are found as two copies of the unit in tandem, the first (a/b) stretching from +5184 to +5425, the second on the opposite strand (c/d) stretching from +5525 to +5729. Also shown in figure 4.9a are the sequence similarities (percent) between the repeat units and the consensus sequence. The alignment indicates that a/b and c/d are diverged from each other to the extent that they show sequence similarity of only 43%. The bovine corticotropin-β-lipotropin precursor gene (bachra - Watanabe et al., 1982) repeat (bachra bases 2653-3920) finds good sequence similarity only with a/b. Baluli1, baluli2 (Duncan, 1987) and the bachra repeats are very similar to a/b and show about 85% sequence similarity. c/d, however, shares good sequence similarity (75-80%) with

Figure 4.9 Similarity of SS1 sequences with ruminant repeat family. (a) shows the DNA sequence of the ovine BLG gene (both strands), over a region of 850 bp, which contains sequences that share sequence similarity with characterised bovine (Watanabe et al., 1982; Skowronski et al., 1984; Duncan, 1987), ovine (Powell et al., 1983) and goat (Spence et al., 1985) repeats. The sequences have been aligned to BLG gene sequence where best match is found. Sequence of the previously characterised repeats is shown above, or below, the SS1 sequence. The boxed region shares similarity with the derived 120 bp basic repeat unit, arrows indicating 5' to 3' direction. The percentage similarity between each of the four repeat units in the SS1 sequence, to the consensus repeat sequence (Watanabe et al., 1982) is shown. The repeat units are labelled a-d. Hatched boxes indicate the *EcoRI* (gaattc) and *BamHI* (ggatcc) sites. (b) shows a map of the SS1, some restriction sites and the seven exons are shown. The region containing the repeats is shown enlarged and positions of the repeat units and their orientations, are shown. The kb scale refers to the upper figure only.


a
a (50.0%)
b (58.3%)
c (73.7%)
d (41.6%)
BLG
balulil
balulil2
bachra
bsat1709
sheep
goat
b
1 kbp
a
b
c
d

a sheep (Powell et al., 1983) and a goat (Spence et al., 1985) repeat (see figure legend for details and references). Figure 4.9b shows a schematic representation of the four 120 bp repeat units and their relationship to each other in the context of the BLG gene.

These data suggest that an ancestral 120 bp repeat unit (which may have had a sequence most similar to c, suggested by its greatest similarity to the derived consensus) duplicated to give the larger, approximately 240 bp unit. The dimeric unit may have arisen independently in a/b and c/d, as suggested by the fact that whereas a/b contains two copies of the repeat in tandem with no sequence separating the two 120 bp units, c/d overlap each other by about 12 bp. There is no evidence to indicate whether a/b-c/d themselves form a yet larger repeat unit or whether they have always been duplicated separately. Certainly none of the other sequenced repeat regions show an arrangement similar to that in the BLG gene.

Sequence comparisons showed that the bovine 1.709 satellite contains the repeat sequences described above. Most satellite DNAs are composed of short repeat sequences. The 1.709 satellite contains little ordered structure. The presence of these Alu-like repeat sequences in satellite DNAs is also unusual. Three copies of the 120 bp repeat unit are present in the 1.709 satellite. Alignment with the BLG repeats showed that two of these are most homologous to repeat unit c, the third to repeat unit b.

Characterisation of the ovine BLG gene shows that at least three repeat regions are present. One of these has been sequenced and shows that over a region of about 700 bp four repeat units of 120 bp, are present as two sets of "dimers". Extensive analysis of this repeat has been carried out by Watanabe et al. (1982) and Skowronski et al., (1984). The repeat units in the BLG gene all show sequence

similarity with a consensus unit derived by these authors.

4.6 TRANSCRIPTIONAL REGULATION

In general, sequences 5' of the transcription initiation site control rates of transcription and temporal and tissue-specific expression. Other sequences, possibly present within introns or in 3' flanking sequences, may also contain regulatory sequences (see section 4.1). Results presented in chapter 3 show that the BLG gene is expressed at low levels in virgin ewes but mRNA levels increase sharply between days 90 and 100 of pregnancy and continue to increase into early lactation. Transgenic mice containing SS1 *SalI* (3.9 kb of 5' flanking DNA, 7.3 kb of 3' flanking DNA) and *SalI/XbaI* (3.9 kb of 5' flanking DNA, 1.9 kb of 3' flanking DNA) fragments express the BLG gene only in the mammary gland, producing BLG in mouse milk (Simons et al. 1987). Two lines transgenic for SS1 *SalI/XbaI* show a time course of expression similar to that of the BLG gene in sheep (S. Harris, unpublished data). A construct containing no more than the region sequenced (figure 4.4) has been injected into mice. Transgenic mice containing this DNA express the BLG gene in the mammary gland only (Harris et al., unpublished data). Thus these sequences are sufficient for expression of the ovine BLG gene. Furthermore, sequences contained within the 3.9 kb region from the *SalI* site to a *PvuII* site at +30 are sufficient to drive mammary-specific expression of a liver-specific gene in transgenic mice (A. L. Archibald, M. McClenaghan, J. P. Simons and A. J. Clark, unpublished data). This suggests that sequences within -811 to +30 are sufficient for tissue-specific expression of the ovine BLG gene.

Deletional and mutational analysis of transcriptional control sequences have led to localisation of important sequences in a number of genes. Much of the eukaryotic gene expression work has been done with viral genes, in particular SV40, polyoma and Adenovirus genes. Much work has also been carried out with the immunoglobulin genes, the globin genes and with many other genes. The current state of knowledge is summarised in the introduction to this chapter and in the reviews cited. Put simply, transcriptional control involves DNA-binding proteins which recognise short, specific sequences and bind to these sequences. The bound proteins are able to influence transcription by RNA polymerase II presumably by protein/protein interactions with the TATA-factor and/or RNA polymerase II (and/or other general transcription factors(?)). It is also possible that changes in DNA conformation upon binding by transcription factors influences transcription.

Techniques such as DNA footprinting, methylation interference and UV-crosslinking have enabled contacts between transcription factors and DNA to be analysed and binding sequences for many factors have been pinpointed exactly. This makes it possible to search for possible binding sites in other gene sequences. The problems associated with such analysis have been outlined in the introduction to this chapter. I will add here that the sequence motifs searched for are consensus sequences derived from a number of binding sites in one or more genes. The sequences are known to show some "redundancy" although some bases may be absolutely essential. They are short so there is a possibility of their presence due to chance. The position of bound transcription factors relative to other factors is likely to be very important for interactions between different transcription complex proteins. Indeed, in a number of genes altering the spacing between elements can reduce transcription. This has been shown for the TATA box which is always present about 30 bp upstream of the transcription start site (Benoist and Chambon, 1981;

Mathis and Chambon, 1981). It is not, therefore, possible to make many predictions with regard to these considerations. Nevertheless, these comparisons can offer information on transcriptional regulation and may indicate experiments which may aid in dissecting promoter action.

The searches were carried out using the UWGCG program FIND (Devereux et al., 1984) which searches for short sequences within a larger sequence and can allow some redundancy (mismatches) between the reference sequence element and the sequence being searched. The search was done on the 5'-most 1000 bp of the BLG gene sequenced, which contain exon I and about 50 bp of intron I as well as the 5' flanking sequences.

Figure 4.10 shows this BLG gene sequence, together with the putative transcription factor binding sites. The mismatch that was allowed depended on the size of the sequence element; for the shortest sequences, such as the SP1 binding site (5'GGGCGG3'), no mismatch was allowed. Table 4.2 lists the sequence elements used in the search, the number of matched sequences found (with the redundancy allowed) and the probability of obtaining such a sequence randomly. A program developed by A. Springbett (IAPGR, Edinburgh) was used to determine probabilities of the presence of random sites for each sequence element. The program utilises the base composition of the sequence being searched. This may be important especially since many of the transcription factor binding sites are G+C-rich. The mammalian genome, on the other hand contains only about 40% G+C (50% would be expected). The ovine BLG gene is unusual in that it has a G+C content of 60% throughout the gene (see below). Thus a greater number of randomly generated putative sites might be expected. The probability generated from this program can be multiplied by 1000 to give an estimate of the number of sites expected in this sequence (since $n + (n - 1) + (n - 2) + \ldots + (n - (j+2)) + (n - (j - 1)) + (n - j)$ overlapping sequences of size j

Table 4.2: Search for *cis*-acting transcriptional control elements

Consensus DNA binding sites are shown. These transcription factor binding sites were compiled from Jones et al. (1988), Chodosh et al. (1988), Nowock et al. (1985), Stuart et al. (1985), Kumar and Chambon (1988), Pelham (1985), Beato et al. (1987). The DNA binding site was used to search for similar sequence motifs in SS1 5'-most 1000 bp sequenced (see figure 4.10) using the UWGCG program FIND (Devereux et al., 1984), allowing up to 4 mismatches. The first row shows the number of finds, the figures in brackets indicate finds on the reverse strand. Greater than 20 finds have not been listed. The second row shows the calculated probability of a find being statistically significant, allowing 0, 1, 2, 3 or 4 mismatches (see text for details). Probabilities lower than 9.9×10^{-6} are not displayed. K = G or T; M = A or C; N = A, G, C, or T; R = A or G; W = A or T; Y = C or T.

Table 4.2: Search for *cis*-acting transcriptional control elements.

Transcription factor	DNA binding site	Number of mismatches 0	1	2	3	4
Octamer	ATTTGCAT	0 (0) 1x10-5	0 (0) 3x10-4	3 (0) 4x10-3	20 (18) 3x10-2	 1x10-1
E2F	TTTCGCGC	0 (0) 2x10-5	0 (0) 5x10-4	2 (0) 5x10-3	11 (13) 3x10-2	 1x10-1
AP1	TKAGTCA	0 (0) 7x10-5	1 (3) 2x10-3	10 (12) 2x10-2	 9x10-2	 3x10-1
AP2	CCCCAGGC	0 (0) 4x10-5	4 (4) 9x10-4	16 (18) 9x10-3	 1x10-2	 1x10-1
AP3	GGGTGTGGAAAG	0 (0)	0 (0) 3x10-6	0 (0) 5x10-5	1 (2) 5x10-4	2 (1) 3x10-3
AP4	CAGCTGTGG	0 (0) 6x10-6	0 (1) 2x10-4	4 (3) 2x10-3	19 (20) 1x10-2	 6x10-2
AP5	CTGTGCNNGGCAAC	0 (0)	1 (0)	0 (0)	0 (0) 4x10-6	6 (6) 6x10-5
PEA2	GACCGCA	0 (0) 1x10-4	2 (3) 2x10-3	13 (14) 2x10-2	 9x10-2	 3x10-1
EF.C	GTTGCNNGGCAAC	0 (0) 1x10-6	0 (0) 1x10-5	0 (0) 2x10-4	2 (1) 2x10-3	4 (8) 1x10-2
E4EF2	TGGGAATT	0 (0) 9x10-6	0 (0) 2x10-4	5 (3) 3x10-3	25 (19) 2x10-2	 1x10-1
E4TF1	GGAAGTG	2 (0) 7x10-5	2 (3) 2x10-3	14 (15) 1x10-2	 8x10-2	 3x10-1
SP1	GGGCGG	1 (0)	9 (9) 2x10-3	 3x10-3	 2x10-1	 5x10-1
MLTF	GGCCACGTGACC	0 (0)	0 (0) 5x10-6	0 (0) 8x10-5	0 (2) 8x10-4	8 (2) 5x10-3
CP1	YN_6RRCCAATCCANYK	0 (0)	0 (0) 3x10-6	0 (0) 5x10-5	1 (0) 5x10-4	10 (8) 3x10-3
CP2	YAGYN_3RRCCAATCN_3R	0 (0)	0 (0) 8x10-6	0 (2) 1x10-4	4 (0) 1x10-3	8 (7) 6x10-3
NF1	NTTGGCN_5GCCAAN	0 (0) 2x10-6	0 (0) 4x10-5	1 (0) 7x10-4	4 (0) 5x10-3	20 (0) 3x10-2
TGGCA	YNTWNYNYTGGMAN$_3$W GCCAANNYN$_5$Y	0 (0)	0 (0)	0 (0)	0 (0) 3x10-6	1 (0) 3x10-5
MRE	CTNTGCRNCGGCCC	0 (0)	0 (0) 1x10-5	0 (0) 2x10-4	2 (0) 1x10-3	5 (2) 9x10-3
HSE	NNCNNGAANNTTCNNGNN	0 (0) 1x10-5	1 (0) 3x10-4	1 (0) 4x10-3	4 (0) 3x10-2	(0) 1x10-1
CREB	KWCGTCA	0 (0) 2x10-4	1 (2) 4x10-3	20 (19) 3x10-2	 1x10-1	 4x10-1
ERE	GGTCANNNTGACC	0 (0) 1x10-6	0 (0) 4x10-5	2 (0) 5x10-4	7 (0) 4x10-3	20 (0) 2x10-2
GRE/PRE	GGTACANNNTGTYCT	0 (0)	0 (0) 2x10-6	0 (0) 3x10-5	0 (0) 4x10-4	5 (7) 3x10-3

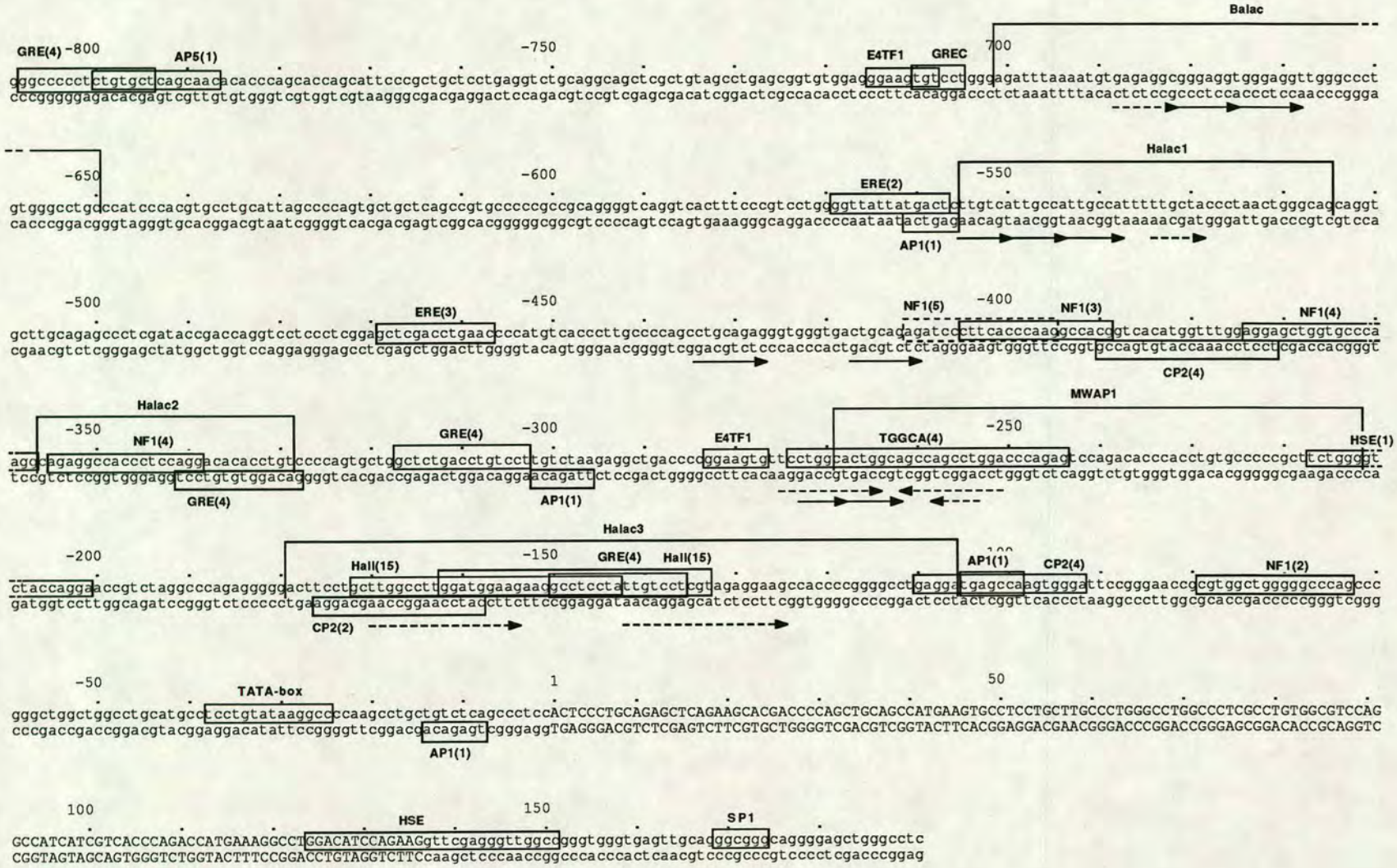
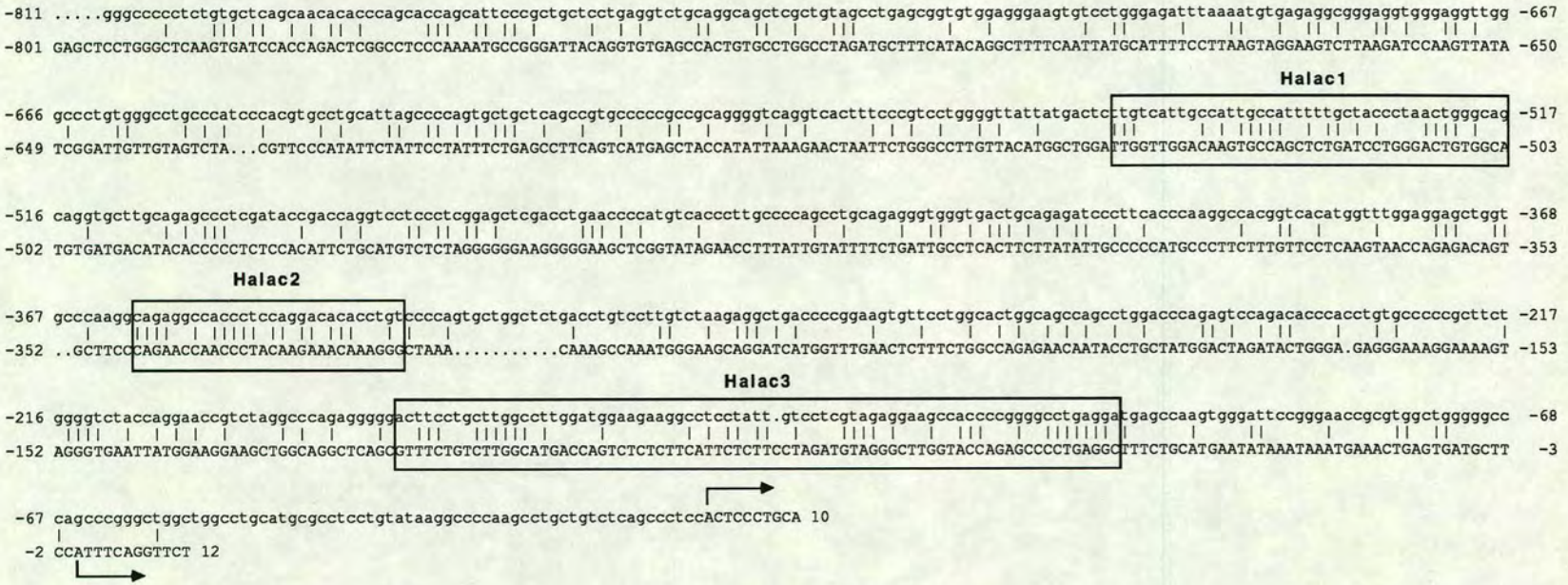

Balac
GRE(4)
-800
AP5(1)
-750
E4TF1
GREC
700
Halac1
-650
-600
ERE(2)
-550
AP1(1)
-500
ERE(3)
-450
NF1(5)
-400
NF1(3)
NF1(4)
CP2(4)
Halac2
-350
NF1(4)
GRE(4)
-300
E4TF1
MWAP1
TGGCA(4)
-250
HSE(1)
GRE(4)
AP1(1)
Halac3
-200
Hall(15)
-150
GRE(4)
Hall(15)
AP1(1)
CP2(4)
NF1(2)
CP2(2)
-50
TATA-box
1
50
AP1(1)
100
HSE
150
SP1

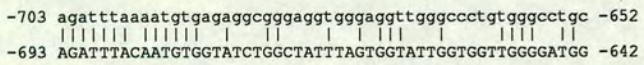
Figure 4.10 Putative *cis*-acting transcriptional control elments in the ovine BLG gene 5' flanking sequences. The 5'-most 1000 bp sequenced are shown. These contain about 810 bp of 5' flanking, exon I and about 50 bp of intron 1 sequences. The numbering shows distance from the transcriptional start site. Boxed sequences show matches to the consensus binding sites of the transcription factor named above (or below) the box. Numbers in brackets show the number of mismatches from the consensus DNA binding site (absence of a number means 0 mismatch). The matches were found, as described in table 4.2 and in the text. The bracketted regions show sequence similarity of the 5' flanking sequences to other milk protein gene 5' flanking sequences (described in figure 4.11). Balac is bovine α-lactalbumin, Halac1-3 are human α-lactalbumin and MWAP1 is mouse WAP sequences which share some sequence similarity with the BLG gene 5' flanking sequences. Arrows indicate the extent of direct, or inverted, repeats, broken lines indicating imperfect repeats.

Figure 4.11 Comparison of BLG gene 5' flanking sequences with other milk protein gene promoters. The figure shows sequence alignments performed using the UWGCG program GAP (Devereux et al., 1984). (a) SS1 gene 5' flanking sequences were compared with the human α-lactalbumin gene 5' flanking sequences (Hall et al., 1987). Regions of relatively high sequences similarity (50 to 65%) are boxed and have been indicated in figure 4.10. (b) SS1 gene 5' flanking sequences were compared with the bovine α-lactalbumin gene 5' flanking sequences (Vilotte et al., 1987). One region of relatively high sequence similarity (51%) was found. (c) SS1 gene 5' flanking sequences were compared with the mouse WAP gene 5' flanking sequences (Campbell et al., 1984). One region of relatively high sequence similarity (54%) was found. No such sequence similarities were noted when SS1 sequences were compared with rat and bovine casein gene 5' flanking sequences. In figure 4.11 vertical lines denote matches; the numbers refer to distance from the transcription start site. In each case the top sequence is SS1.

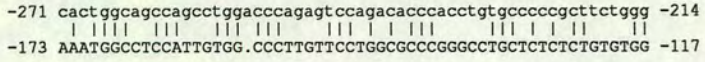
a. Ovine BLG gene v Human a-lac gene



b. Ovine BLG gene v Bovine a-lac gene



c. Ovine BLG gene v Murine WAP gene



present in a sequence of size n). Thus in this case a probability of 10^{-3}, or greater, means that any match found is not statistically significant, but a probability of 10^{-4}, or less, is statistically significant. Due to other considerations, discussed in the introduction to this chapter and in this section (above), sites not statistically significant could nevertheless be important transcriptional regulatory sites. A similar analysis was done on the reversed sequence of each DNA binding site, as many transcription factors can apparently act in an orientation-independent manner.

4.6.1 The possible significance of found matches

Dissection of gene promoters has suggested that promoter function can require many transcription factors to bind to the promoter. Thus, many proteins could be involved in initiation of transcription of a gene (for example, see the review by Jones et al., 1988). It is not clear how many such proteins are present in any cell or which cells contain the same proteins. The search for control elements was therefore performed on *cis*-acting elements which have been well characterised and which appear to act in a number of cell types. These sequences are known to be bound by transcription factors, some of which have been cloned and others which have been purified to homogeneity. All the proteins are transcriptional activators, although some are also known to act as negative regulators. They are shown in table 4.2.

The table shows the number of sites generated when four or less mismatches are allowed. It also shows the probability of a random occurrence of a match in a sequence having the base composition seen in the BLG gene (see above). I have disregarded sequence positions which have a probability of greater than 2×10^{-4} of presence by chance, except when other conditions have been imposed.

These transcription factors listed have been well characterised. In

particular, many steroid hormone receptors (including the oestrogen, glucocorticoid and progesterone receptors) have been cloned (Green and Chambon, 1988; Evans, 1988). Their sequence requirements are well defined. It has been shown that the glucocorticoid and progesterone receptors bind to similar or overlapping recognition sequences (von der Ahe et al., 1985, 1986; Strahle et al., 1987; see also section 4.1, this thesis for discussion). Comparison of 22 glucocorticoid receptor elements suggests an almost absolute requirement for the core (GREC) sequence (TGT(C/T)CT). The other half of the consensus sequence is much less well conserved. For this reason 3 out of the 5 statistically non-significant (4 mismatches in this case) motifs generated by the search have been shown in figure 4.10. Two of these sites contain TGTCCT whilst the third differs in one position from the core consensus (this mismatch has been observed in functional GREs (Beato et al., 1987)). Other GRECs have also been shown in the figure. The GREs, if functional, may be bound by the progesterone receptor as outlined in section 4.1 and above.

Martinez et al. (1987), Klein-Hitpass et al. (1988) and Kumar and Chambon (1988) have characterised oestrogen receptor element (ERE) mutants which define some non-mutable bases. Of the matches to the ERE found in the BLG gene, only two contained all the known essential bases.

A number of CCAAT-binding proteins have been described (Jones et al., 1987; Dorn et al., 1987; Chodosh et al., 1988; Santoro et al., 1988). The proteins bind to similar but distinct sequence elements. The trinucleotide CCA is always part of the bound sequence. Dorn et al. (1987) showed that mutation of any of these three nucleotides led to almost total abolition of activity from a NF1 binding element. Thus, matches lacking this trinucleotide were disregarded.

Two good matches to the heat shock elements (HSE) were obtained. This is likely to be a fortuitous match, however, since there is no evidence for heat shock

inducibility of milk protein genes or for the BLG gene. Nor is it possible to envisage a role for such a gene in the heat shock response.

A very good putative AP5 binding site is present at the very 5' end of these sequences. AP5 is important for the function of the SV40 enhancer in HeLa cells (see Jones et al., 1988). Mutation of the AP5 recognition site greatly decreases enhancer action.

Figure 4.10 also shows regions of sequence similarity between the BLG gene 5' sequences and α-lactalbumin, WAP and casein gene 5' sequences. No real sequence similarity was seen between the BLG gene and any of the casein genes. Short regions of sequence similarity between the BLG gene and the mouse WAP gene (figure 4.11c) and between the BLG and bovine α-lactalbumin genes (figure 4.11b) were found using the UWGCG program GAP (Devereux et al., 1984). Comparison of the human α-lactalbumin and ovine BLG genes (figure 4.11a) showed three regions of similarity. These regions have been indicated in figure 4.10.

In addition, some short, direct, or inverted, repeats are present in the 5' sequences and are indicated in figure 4.10. Many transcription factor binding sites contain short repeats, often separated by a few nucleotides. This gives a dyad symmetry, for example in the GRE and ERE; here the receptors apparently interact with their recognition sites as dimers (Kumar and Chambon, 1988; Tsai et al., 1988) (for example, the heat shock and NF1 elements also have dyad symmetry). The short, direct and inverted repeats may indicate binding sites for a dimeric protein. The larger repeats may indicate multiple factor binding sites in tandem. Enhancers often contain multiple factor binding elements, allowing cooperative binding of transcription elements (see section 4.1). Sequences similar to these repeats were found in other milk protein gene 5' flanking sequences but none as tandem direct, or inverted repeat, as seen in the BLG gene sequence.

Sequences downstream of about -410 contain much of the putative transcriptional element structure and may therefore be the sequences important for tissue-specific and temporally-regulated expression of the BLG gene. Sequences upstream of this region could also be involved in regulation. Certainly some possible binding elements are present. For example, the two possible EREs and the very good AP5 match are present in the region 5' of -410. In the sequences from -410 to the transcription start site the presence of four domains can be postulated. The first stretches from -410 to -330 and contains four putative CCAAT-like elements. The distance between the central CCAA in the first, third and fourth NF1 sites in this region (5' to 3') and the CP2 site (on the other strand) is 19 bp (from the first NF1 site to the CP2 site), 20 bp (from CP2 to the third NF1 site) and 20 bp (from the third to the fourth NF1 sites). This would mean that the CCAAT elements are separated by two exact turns of the DNA helix in two cases and almost exactly two turns separate the first NF1 site from the CP2 site CCAA. Thus, the four bound proteins would be able to interact with each other by being bound in tandem.

The second region also contains a CCAAT-binding protein site (from -280 to -245). This is a TGGCA-binding protein site which differs from the consensus by a switch of two bases from the TGGCA-binding protein core sequence. The TGGCA-binding protein core TGGCANNNTGCCA is here TGGCANNNAGCCT. The short repeats centred here suggest that this may be a binding site for an NF1/TGGCA-like transcription factor. Lubon and Hennighausen (1988) demonstrated DNaseI protection of a similar sequence motif situated between -125 and -101 of the rat α-lactalbumin gene. They also showed the presence of this sequence as part of a highly conserved sequence in four whey protein genes (human and rat α-lactalbumin and murine and rat WAP genes). In these genes, however, the sequence element is

present at around -120 and -140, whereas in the BLG gene it is found at around -280. Furthermore, they showed that purified HeLa cell NF1 can bind to this sequence in the rat α-lactalbumin gene.

Lubon and Hennighausen (1987) showed that four high affinity complexes are found between -175 and -88 of the mouse WAP gene. Comparison of the mouse WAP gene 5' sequences with the ovine BLG gene using GAP revealed sequence similarity over a region of the mouse WAP gene stretching from -173 to -117 (figure 4.11). A strongly protected complex is formed on the mouse WAP gene between -177 to -160 and a weaker one from -155 to -135. Another strong complex stretches from -135 to about -95. The strong complex at around -170 contains a sequence motif similar to the TGGCA-binding protein site, although not as good a match to the consensus site as the one present in the BLG gene.

Hall et al. (1987) described a highly conserved region present at -140 to -110 in seven milk protein genes of cow, guinea pig, man and rat (see also Laird et al., 1988). Lubon and Hennighausen (1988) subsequently showed that part of this sequence contains the TGGCA-like sequence mentioned above. No good match with the entire Hall et al. (1987) sequence was seen for the ovine BLG gene, the best sequence similarity not exceeding 50-55%. None of these matches contained a TGGCA motif. Figure 4.10 shows two of the best matches to the Hall et al. (1987) consensus sequence. These are shown as they are present in a region similar to that observed in other milk protein genes. Hall et al. (1987) and Laird et al. (1988) showed that the other milk protein genes share a sequence similarity of 70-85% in this region. Like the ovine BLG gene the murine and rat WAP genes do not show good sequence similarity in this region. The BLG gene does, however, share some sequence similarity with human α-lactalbumin gene sequences from just downstream of this conserved sequence (figures 4.10 and 4.11).

The 3' end of this region of the BLG gene contains a "classic" CCAAT-box, often present in this region of the promoter in RNA polymerase II genes (Breathnach and Chambon, 1981). The fourth region is a TATA-box, which is present in most RNA polymerase II-transcribed genes (Breathnach and Chambon, 1981) at around -30 (in higher eukaryotes).

Although it seems clear from the unpublished results of Harris et al. and of Archibald et al. that 810 bp of 5' sequences are sufficient for correct expression of the ovine BLG gene, it is possible that other sequences present in introns or in the 3' flanking regions are involved in control of expression. A search using the sequence elements described above, yielded no apparent important regions or significant matches, with the exception of a TGGCA-binding protein site (3 mismatches) present on the (-) strand, at +5346.

4.7 BASE COMPOSITION OF THE OVINE BLG GENE

It was clear from the BLG gene sequences that it has a high G+C content. It is, therefore, possible that the BLG gene contains a CpG island, as described in chapter 3. Analysis of the methylation status of the BLG gene has been initiated and preliminary results, presented in chapter 3, suggest that there are methylation differences between expressing (mammary) and non-expressing (liver) tissue DNAs. To determine the nature of the BLG gene sequence and to see whether a CpG island is present, the base composition has been analysed. The 7389 bases show an overall G+C content of 59.6% (29.8% G and 29.8% C), with an A+T content of 40.4%. This is an exact reverse of the situation normally found in vertebrate

genomes. Figure 4.12 shows the base composition (percentage) across the length of the BLG gene. This shows that the G and C content is almost always greater than normal and the A and T base composition is therefore mostly lower than expected. At only two regions in the gene (at the 5' end of intron II and in the sequences 3' of the final exon) does the A and T composition really reach and even exceed its expected levels. The ovine BLG gene therefore, appears to form part of an unusual G+C-rich DNA domain.

CpG islands are G+C-rich regions but are usually no more than 1-2 kb long. The BLG gene is unusual in having a 7 kb (or longer?) stretch. Lindsay and Bird (1987) calculated the distribution of restriction endonuclease cleavage sites for some enzymes which contain one, or two, CpG dinucleotides in their recognition sequences, in bulk mammalian DNA and in CpG islands. By their criteria 89% of all *NotI* (GCGGCCGC) sites should be in CpG islands (0.12 sites per island; rare even in CpG islands), 74% of *EagI* (CGGCCG), *SacII* (CCGCGG) and *BssHII* (GCGCGC) sites should be present in CpG islands (1.2 sites per island). Restriction mapping SS1 with these restriction enzymes showed the presence of two *NotI* (and *EagI*) sites within 10 kb (one in intron 1, at +346), the second about 2.0 kb 3' of the *XbaI* site. A further *EagI* site is present 1.2 kb 3' of *XbaI*. A *SacII* site is present close to (and 5' of) the *EcoRI* site 3' of *XbaI*. Sites for *NarI* and *SmaI*, other CpG enzymes which are less rare in bulk DNA, but nevertheless over-represented in CpG islands, are present 3.5 and 1.4 kb upstream of the *SphI* site, respectively. This suggests that the G+C-rich domain of the BLG gene probably extends over a region of greater than 13 kb. This would be highly unusual for a CpG island.

In any case, to determine whether the BLG gene also forms a CpG island the CG content was calculated. The dinucleotide composition of the gene is shown in table 4.3. 193 CG dinucleotides are present whereas 528 GC dinucleotides are found.

Table 4.3 BLG gene base composition

*****

A: 1457 (19.7%)	C: 2200 (29.8%)	G: 2201 (29.8%)	T: 1531 (20.7%)

Other: 0

Total: 7389

*****

GG: 847 (11.46%)	GA: 460 (6.22%)	GT: 366 (4.95%)	GC: 528 (7.15%)
AG: 554 (7.50%)	AA: 294 (3.98%)	AT: 243 (3.29%)	AC: 365 (4.94%)
TG: 606 (8.20%)	TA: 147 (1.99%)	TT: 324 (4.39%)	TC: 454 (6.15%)
CG: 193 (2.61%)	CA: 556 (7.53%)	CT: 598 (8.09%)	CC: 853 (11.55%)

Other: 0

Total: 7388

*****

	Whole gene		5' sequence	
	Expected	Actual	Expected	Actual
G+C	40%	59.6%	40%	63.9%
A+T	60%	40.4%	60%	36.1%
CG	665	193	90	31
GC	665	528	90	95
Ratio GC/CG	1	2.74	1	3.06
TG+CA	887	1162	120	165
GT+AC	887	731	120	92
Ratio TG+CA/GT+AC	1	1.59	1	1.79

GC - CG = 335 GC - CG = 64

(TG+CA)-(GT+AC) = 431 (TG+CA)-(GT+AC) = 73

The number of nucleotides and dinucleotides are shown, as well as percentage abundance. If random composition is assumed, then each nucleotide should represent 25% of the total and each dinucleotide should represent 6.25% of the total. Also shown is an analysis of G+C content and CG dinucleotide content of the entire gene and of the 5' flanking sequences. The expected values of G+C and A+T are as quoted by Bird (1986). Expected CG and GC numbers were determined by multiplying the total number of dinucleotide pairs expected (7388) by C and G percentages (taken as 30/100 for each). The expected GC/CG ratio was taken as 1 (for a CpG island - Bird, 1986). The numbers of TG+CA and GT+AC and the TG+CA/GT+AC ratio have been calculated for comparison. GC - CG and (TG+CA) - (GT+AC) numbers are also compared, to test whether the deficiency in CG dinucleotides, relative to GC, is met by an abundance of (TG+CA), relative to (GT+AC) (see Bird, 1986; and the text).

Figure 4.12 Analysis of the base composition and CpG content of SS1 sequences. The figure shows a map of SS1BH and SS1HX regions (see figure 4.3). Exons and some restriction sites have been indicated. The plot below was drawn using the UWGCG program DOTPLOT (Devereux et al., 1984) and shows CpG, GpC, TpG and GpT dinucleotide presence in the SS1 sequence (on both strands). The vertical bars indicate the presence of a dinucleotide at that point in the sequence. The bottom plot was made using the UWGCG program STATPLOT (Devereux et al., 1984). It shows the percent G, A, T and C content of SS1 (window = 100, shift = 3). The dotted lines show the expected percentage of each nucleotide from gross DNA estimates (G+C = 40%, A+T = 60%). The arrows indicate observed mean percentages for each nucleotide in the entire SS1 sequence (top strand only). The nucleotide and selected dinucleotide plots have been aligned to the map of SS1. Thus, the nucleotide and dinucleotide composition at any point in the sequence can be seen directly.

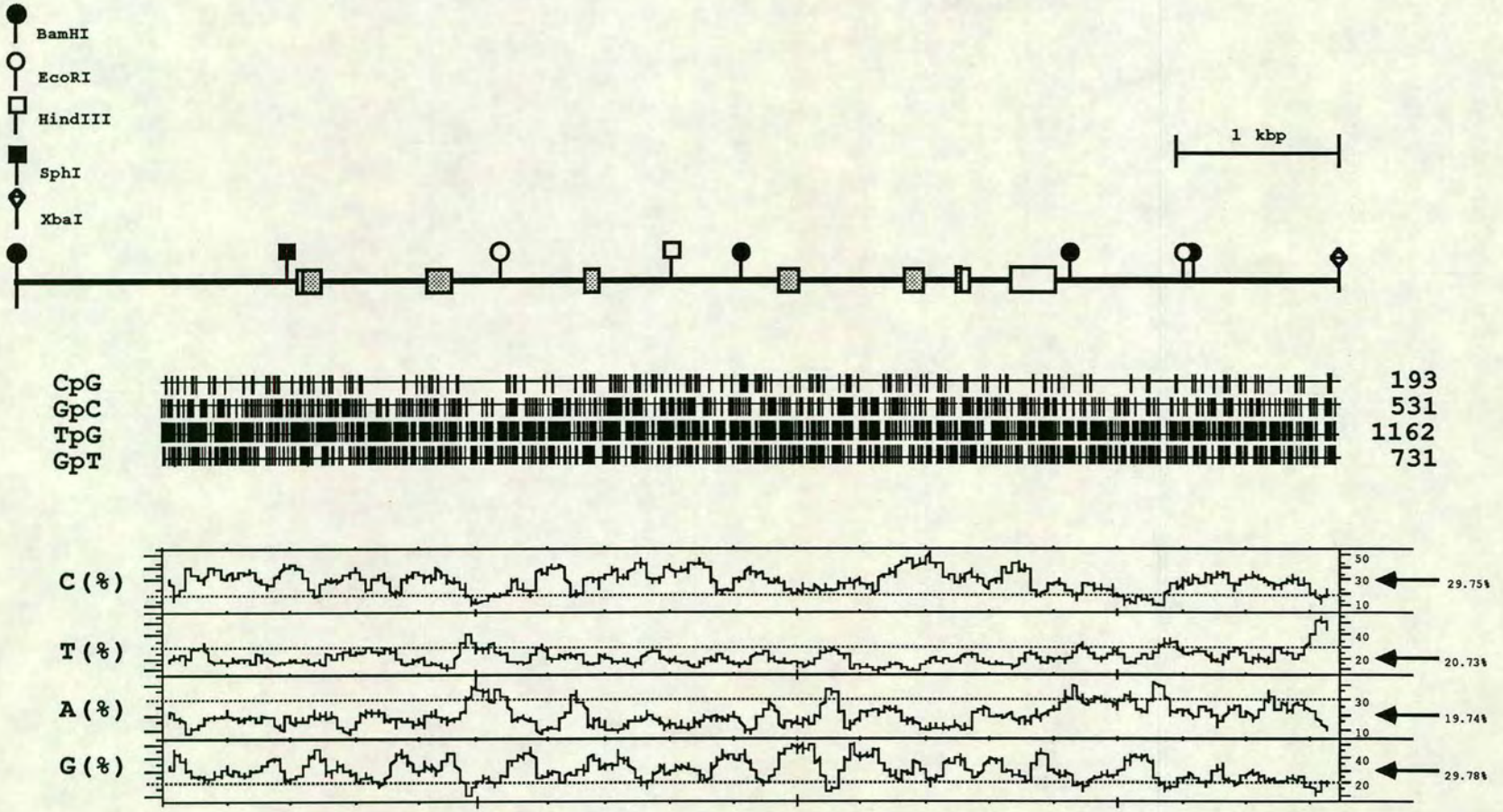
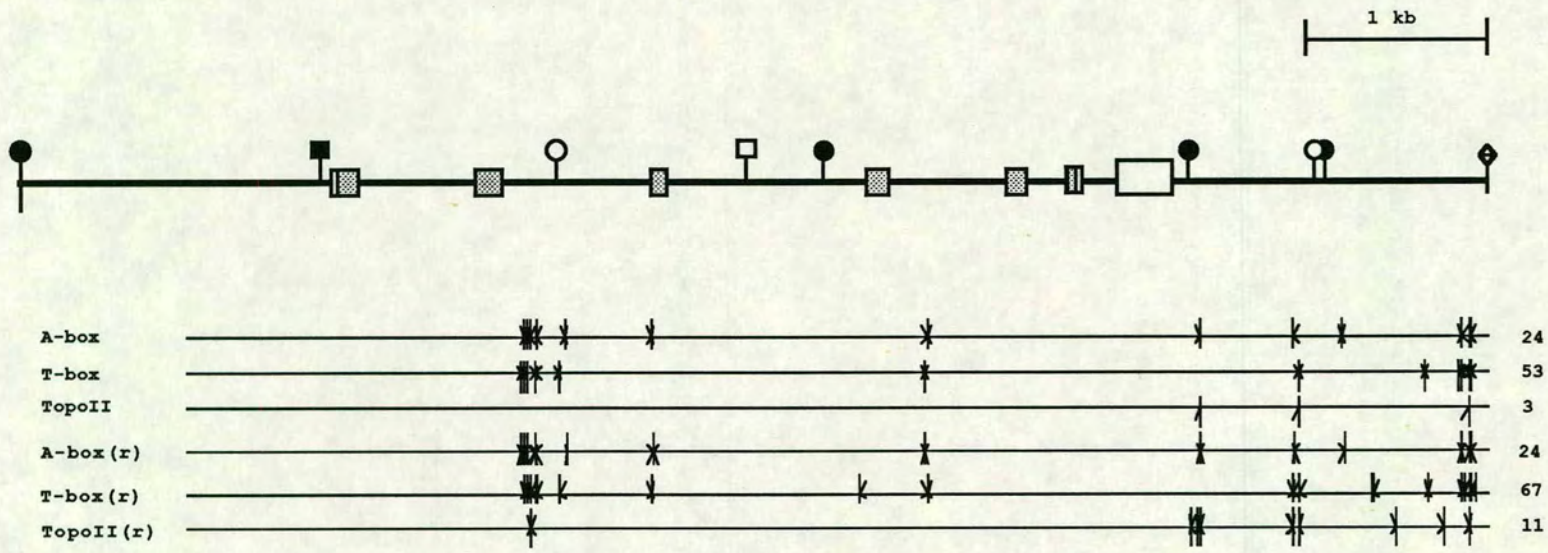

BamHI
EcoRI
HindIII
SphI
XbaI
1 kbp
CpG
GpC
TpG
GpT
193
531
1162
731
C(%)
T(%)
A(%)
G(%)
50
30
10
40
20
30
10
40
20
29.75%
20.73%
19.74%
29.78%

Figure 4.13 Mapping putative Scaffold Attachment Regions from SS1 sequence. This shows a map of the SS1 BLG gene, as in figure 4.12. The sequenced region is represented below as straight lines and the positions of the A-box, T-box and topoisomerase II (topoII) sites and their reversed complements are shown as vertical lines (see text). Mismatches of 3 (≥ 70% sequence similarity) was searched for. The diagonal lines indicate degree of mismatch. More than one site, present close together may be shown by a single line. The UWGCG program MAPPLOT (Devereux et al., 1984) was used to generate the lower plot.


1 kb
A-box
T-box
TopoII
A-box(r)
T-box(r)
TopoII(r)
24
53
3
24
67
11

Although both are less frequent than expected for a sequence of this composition (0.3x0.3x7388 CG and GC pairs expected = 665) CpG islands should contain about equal numbers of CG and GC dinucleotides. This is clearly not the case here (193 and 528 CG and GC pairs present, respectively). The prediction of the hypothesis that 5-methylcytosine is highly mutable to thymine, suggests that the difference in GC and CG numbers should be approximately that between TG+CA and GT+AC (mutation of CG on the other strand will give CA on the top strand so both TG and CA are considered). Comparisons of these dinucleotides showed that TG+CA is present in excess of GT+AC and the number of excess dinucleotides is not greatly different from the difference between GC and CG dinucleotides.

To see whether there was a region which had a high G+C content but also similar CG and GC levels a plot of CG and GC (and TG and GT) was made. This is shown in figure 4.12, together with the exonic arrangement of the BLG gene and the base composition. Each dinucleotide is marked as a vertical line dissecting the horizontal line representing the entire sequence. "TpG" and "GpT" show the dinucleotides TG+CA and GT+AC, respectively. The plot shows that there is no local region of the gene which contains approximately equal numbers of CG and GC dinucleotides. Thus, nowhere in the BLG gene is the requirement for the presence of a CpG island met.

Since CpG islands are usually present at the 5' end of genes table 4.2 also shows the base composition of the first 1000 bp of the gene. This confirms the apparent absence of a CpG island in this region.

The BLG gene appears to be unusual in being part of a very large, G+C-rich domain. It does not appear to contain a CpG island and appear to be deficient in CpG throughout the gene and the difference between numbers of GC and CG dinucleotides is similar to the difference in numbers between TG+CA and GT+AC, suggesting that the gene is methylated. Bernardi and co-workers have suggested that "warm blooded"

vertebrates' DNA consists of a mosaic of large domains of heavy (G+C-rich) and light components (see Bernardi et al., 1985; also see Mouchiroud et al. (1988) for recent references). These domains may be larger than 200 kb in length and can be visualised as buoyant density differences. Furthermore, Bernardi and coworkers suggest that a greater number of genes may be present in the G+C-rich components than in the light ones. Genes in G+C-rich, heavy components contain high G+C (higher than the G+C content of the heavy component), high G+C content at codon third positions (ranging from 43-69% in light component genes and 61-90% in heavy component genes) and high CpG/GpC ratios (approaching closer to 1 with increasingly heavy components). The BLG gene may be present in such a large, G+C-rich, heavy component. The G+C level of the codon third positions is 85% for the BLG gene, in agreement with Bernardi et al.'s findings. However, the CpG/GpC content is far lower than would be predicted from Bernardi's findings. So whilst the G+C content of the gene overall and of codon third positions suggest that the BLG gene is present in a heavy component, the CpG/GpC ratio indicates that it should be in a light component. No clear functions for these light and heavy components have been put forward so their relevance is not understood. Bernardi et al. (1985) have suggested that these components may correlate with the heterogeneity associated with chromosomal banding."Cold-blooded" vertebrates appear to lack such heterogeneity of components, most of their DNA being light.

The two main regions of A+T-richness in the BLG gene sequence both contain very high levels of A+T. One in intron 2 stretches over about 350 bp (1090-1350) and contains a run of 13 Ts. The second region starts 3' of the final exon. The sequence 3' of 6400 is particularly A+T-rich, with runs of Ts (with lengths of 7, 13 and 6) together with a stretch of $(TA)_8$.

Such unusual runs of Ts may be involved in gene regulation. The form of

DNA mostly present in living cells is the B form. There are 10 bases per complete turn of the double helix. Other forms of DNA have been described (at least in vitro). The D and E forms (8 and $7^1/_2$) turns respectively) are taken up by some DNA molecules lacking guanines. Z-DNA forms a left-handed helix (the other forms are right-handed) and can be formed by alternating purine-pyrimidine tracts (for example see Lewin, 1983). Z-DNA has been postulated to play a role in gene regulation due to its different conformation (see Rich et al., 1984).

Runs of As and/or Ts have been described for other genes. For example, the different mouse Major Urinary Protein genes (Al-Shawi et al., 1989) contain variable lengths of As which can be as many as 60, interspersed with a few Cs. This region is present just upstream of the TATA-box and is probably important for transcription. Struhl (1986) showed that a 17-bp (dA-dT) tract is responsible for constitutive expression of the yeast bidirectional *his3* and *pet56* genes. This sequence is present about 110 bp upstream of the *his3* gene transcriptional start site and about 60 bp upstream of the *pet56* gene start site. The region shows nuclease sensitivity which is lost on deletion of the the poly (dA-dT). Struhl (1986) has proposed that these sequences act by excluding nucleosomes. None of the poly (dA-dT) stretches in the BLG gene is present near the transcriptional start site. Nevertheless an "open" domain (nuclease sensitive) may be able to act at some distance from the promoter in any transcriptional activation function.

Scaffold-attached DNA regions (SARs) have been proposed to be involved in formation of domains of chromatin by forming loops which may compartmentalise chromatin (see Gasser and Laemmli (1987) for review and further references). SARs have been mapped to small restriction fragments in *Drosophila* and in rodent cells. Many are found both 5' and 3' of highly expressed genes and have been comapped with enhancer elements upstream of three developmentally regulated genes

in *Drosophila* (Gasser and Laemmli, 1986) and flanking the mouse immunoglobulin κ-gene enhancer (Cockerill and Garrard, 1986). The presence of a SAR can eliminate position effects in transformation experiments, reducing expression level variation, due to different integration sites. Grosveld et al. (1987) found position-independent high levels of expression of the human β-globin gene in transgenic mice when two regions, one in the 5', the other in the 3' flanking sequences, were present. Previous experiments have given variable levels of expression (see Grosveld et al. (1987) for references). These form two DNaseI hypersensitive regions which appear to be erythroid-specific. The authors have speculated that the regions may be SARs. It is therefore, possible that SARs may indeed compartmentalise genes to transcriptionally active domains to create functional complexes for transcriptional regulation. Available evidence, however, suggests that there is little tissue-specificity, with no differences in loop organisation being seen in different cell types in mouse (Cockerill and Garrard, 1986) or Drosophila (Gasser and Laemmli, 1986). If the "locus-activating regions" described by Grosveld et al. (1987) are SARs they differ from previous observations in that they act tissue-specfically. Two major scaffold-associated proteins have been identified, one of which is topoisomerase II. Two sequence motifs have also been identified within SARs. These are the so-called "A-box" (AATAAAYAAA) and the "T-box" (TTWTWTTWTT), present in multiple copies in the SAR regions. The significance of these motifs is not yet known. The topoisomerase II-binding site consensus sequence (GTNWAYATTNATNNG) has also been noted in this region. This sequence motif is present much more often in SARs than in other sequences, forming clusters of potential topoisomerase II recognition sites in SARs.

A search for these sequences was carried out in the BLG gene sequences, using the UWGCG program MAPPLOT (Devereux et al., 1984) which generates a plot

as shown in figure 4.13. Clustering of potential A-box, T-box and topoisomerase II sequences is noticeable in two regions. The first is in intron II, the second is in the 3' flanking sequences. Sequences with greater than 70-80% sequence similarity were searched for. The presence of these sequence motifs may simply be a reflection of the high A+T content, rather than functionally significant. Nevertheless, these are unusual regions in such a G+C-rich domain.

4.8 SUMMARY

The β-lactoglobulin gene is tissue-specific, being expressed in the mammary gland. Its expression is temporally-regulated, showing similar regulation to some ovine milk protein genes and differing, in its expression, from other milk protein genes. To determine how it is expressed and to analyse mechanisms of transcriptional control, the cloning and sequencing of the gene is required. A sheep spleen genomic DNA library, screened with the ovine BLG cDNA yielded four clones which were shown to be very similar, by restriction mapping (A. J. Clark's results; Ali and Clark, 1988). Restriction enzyme digests of sheep genomic DNA suggest that no major rearrangements have occurred in cloning (figure 4.2). These data also indicate that the ovine BLG gene is encoded by a single copy gene. It is clear, however, that one or more BLG-like genes are present in sheep. These may be pseudogenes or they may be functional genes expressed in other tissues. There is no evidence that they encode a BLG gene which is expressed in the mammary gland. The presence of a number of repeats was also found.

DNA sequencing shows that ovine BLG is encoded by a seven exon gene containing the sequence motifs known to be present in most RNA polymerase II genes

(figures 4.3, 4.4 and 4.5; table 4.1), suggesting that SS1 and SS12 (see chapter 5) encode functional BLG genes. Transgenic mice containing SS1 or SS12 sequences secrete BLG into milk, implying that the genes are indeed functional (Simons et al., 1987; chapter 5). The sequence shows that an Alu-like repeat is present 3' of exon 7. Interestingly, this repeat is also present in an unusual bovine satellite (Skowronski et al., 1984).

The 5' flanking sequences contained in the 4 kb *SalI/PvuII* fragment can drive mammary-specific expression of heterologous gene sequences (Archibald et al., unpublished results). Furthermore, sequences from -810 to the *XbaI* site (1.9 kb downstream of exon VII) are sufficient for mammary-specific expression of the ovine BLG gene in transgenic mice. This strongly suggests that sequences in the 5' flanking region from -810 to the transcription start site contain the necessary sequences for correct transcriptional regulation of the gene. Comparison of these sequences with known transcription factor DNA-binding sites has been carried out (figure 4.10).

This comparison shows that there is a region from about -410 to the transcription start site which contains a number of transcription factor binding sites (figure 4.10). It can be postulated that at least four CCAAT-binding protein sites act in tandem, the central CCAA separated by two complete turns of the double-helix. It can be envisaged how factors could bind here in a cooperative manner. The fact that exactly 20 bp separate three of the binding sites (the other distance is 19 bp) seems to suggest a functional significance for this putative domain. A second domain appears to be bound by a similar protein, the TGGCA-binding protein site (Nowock et al., 1985) (-275 to -245) which appears to be closely related, if not identical, to NF1. This site shows a repeat motif around the putative binding site. This region forms part of the best sequence similarity

between the BLG gene and the mouse WAP gene (figure 4.11). The sequences of the WAP gene showing similarity with BLG gene sequences at the TGGCA site are protected from DNaseI digestion by lactating mammary gland nuclear proteins.

A third region (-180 to -90) shares sequence similarity with the human α-lactalbumin gene and contains two putative CP2 binding sites. The CP2 site at around -100 is at the classic CCAAT-box position (see Breathnach and Chambon, 1981). Finally, a TATA-box is present at the expected position (-33) and probably binds the TATA factor, TFIID (BTF1).

The apparent importance of CCAAT-proteins in the control of BLG gene expression is interesting. The CCAAT-proteins are members of a large family of proteins (see section 4.1 for references) involved in transcription activation, but have also been implicated in viral DNA replication (Jones et al., 1987). Although these putative BLG gene domains appear to contain transcriptionally important sequences none have yet been shown to be real. Furthermore, the existence of other important sequences is highly likely, especially the steroid hormone responsive elements. If mammary-specific transcription factors are involved the presence of their recognition elements may not be detected by sequence comparisons, although sequence similarities between different milk protein gene promoters may suggest important sequences. It is also possible that the short repeats at -560 to -530, -690 to 680 and -440 to -410 are factor recognition sites, as many transcription factor binding elements bind to such sequences (as discussed previously). However, the work of Lubon and Hennighausen (1987) showed no mammary-specific binding to the mouse WAP gene (although at least one DNA-protein complex appeared to be weaker with HeLa cell extracts than with mammary extracts). Lubon and Hennighausen (1988) also showed that weaker DNaseI protection over a region containing a TGGCA-like sequence was obtained when HeLa cell nuclear extracts were

used. Purified HeLa cell NF1 specifically binds to this sequence. A number of proteins are known to bind to the "octamer" motif, some of which are tissue-specfic (see Jones et al., 1988). A similar situation could exist in the mammary gland where a site known to be bound by a transcription factor in other cell types may be bound by a mammary-specific factor in the mammary gland (see also section 4.1 for discussion of tissue-specficity of transcription factors).

Further elucidation of transcriptional control will come from production of transgenic mice with constructs having further deletion of 5' sequences, as well as the use of nuclear extracts for actual mapping of DNA-protein complexes (S. Harris, work in progress).

Prolactin (and also placental lactogen and growth hormone) controls milk protein gene expression (see chapters 1 and 3). It is not known how these hormones act at the transcriptional level. They bind receptors on the cell surface and probably act via secondary messengers. Modification of one or more transcription factors may "activate" them to forms capable of transcriptional regulation (see section 4.1). Cyclic AMP (cAMP) (a secondary messenger whose levels are dependent on hormones being bound to cell surface receptors and activating adenylate cyclase) responsive elements have been described. Although it is not known by which mechanism(s) cAMP acts on transcription factors (it is known to activate protein kinase C) deletion studies and DNaseI protection have mapped a cAMP responsive element which contains two CGTCA motifs. Mutations in either motif drastically reduce cAMP-dependent expression of the vasoactive intestinal peptide gene (Fink et al., 1988). No such sequence element is present in the BLG gene 5' sequences presented here. However, it seems that cAMP-regulated transcription factors may form a family of related proteins which includes ATF and AP1. ATF and AP1 have been shown to be immunologically related and to bind similar, although distinct, sequence

elements (Hai et al., 1988). Possible AP1-binding sites are present in the BLG gene 5' sequences (figure 4.10). The steroid hormone receptors form a large family of transcriptional regulators that directly bind DNA. Putative GREs and EREs are shown in figure 4.10 and these may be functional.

Large stretches of DNA, distant from the transcriptional start, may be involved in interactions which regulate expression. Thus, enhancers may be present far upstream or downstream of a gene. Other sequences which have been postulated to be involved in gene regulation include SARs. They may act by bringing transcriptionally active genes (or domains) together, forming compartments. Clear evidence of the involvement of these scaffold regions in gene regulation has not yet been forthcoming.

Other measurable gross features of active genes include nuclease sensitivity and CpG methylation. DNaseI hypersensitive sites have been mapped in a number of genes (for example, the chicken globin genes (Stalder et al., 1980; Larsen and Weintraub, 1982), rat albumin and α-fetoprotein genes (Nahon et al., 1987) and the MHC class II genes (see Peterlin et al., 1987). DNaseI hypersensitive sites are often absent from non-expressing genes and appear on gene activation. Thus, tissue-specfic patterns of hypersensitivity are found (for example Turcotte et al., 1986). Sensitivity is found to extend far 5' and 3' of the transcription unit. Thus, DNaseI hypersensitivity can suggest the presence of actively transcribed genes and indicate the stage of development at which induction of the gene occurs due to appearance of hypersensitivity on gene activation (see Igo-Kemenes et al. (1982) for review). The functional significance of hypersensitive sites is not clear. They may be "open" domains which are more accessible to transcriptional regulators (Stalder et al., 1980). Determination of BLG gene hypersensitive sites may help in determining sequences important for the control of its expression. For example, Wu

(1980) observed the presence of DNaseI hypersensitive sites in *Drosophila* heat shock genes close to the 5' end of the coding sequences.

CpG methylation has been found to correlate with gene activity in many cases. Thus housekeeping genes are undermethylated (in the promoter region), tissue-specific genes are often undermethylated in tissues in which they are expressed and methylated in non-expressing tissues (Razin and Cedar, 1984). Although it is not clear whether DNA methylation in these regions is a cause or an effect of expression, tissue-specific patterns of methylation have often been observed for certain CpG dinucleotide pairs (generally at the 5' end). Thus in the rat γ-casein gene certain CpGs are not methylated in the expressing mammary gland but are methylated in the liver (Johnson et al., 1983). Preliminary work with the ovine BLG gene suggests that there are differences in methylation between mammary and liver DNA.

Genomic cloning and DNA sequencing of the ovine BLG gene has suggested some approaches to the further study and dissection of milk protein gene expression. As one of a few tissues which undergoes a major part of its development after birth, under the control of manipulable signals (i.e. pregnancy) study of the induction of such genes, particularly at the level of transcription, is very interesting. In this chapter I have analysed the ovine BLG gene sequences for the presence of potentially important control sequences and indicated directions in which work is progressing. In particular, as mentioned above, further work to make transgenic mice containing BLG gene sequences with greater 5' deletions, are underway and will help to pinpoint important sequences functionally. Mammary gland tissue obtained during the time course (chapter 3) can also be used for analysing DNaseI hypersensitivity and DNA-protein interactions.

Chapter 5. CHARACTERISATION OF THE GENE ENCODING OVINE BLG-A, EXPRESSION IN TRANSGENIC MICE AND POLYMORPHISM ANALYSIS

Two variants of ovine BLG have been described, termed BLG-A and BLG-B, separable by charge (Bell and McKenzie, 1967). Amino-acid sequencing has indicated that a tyrosine residue in BLG-A is replaced by a histidine in BLG-B (Kolde and Braunitzer, 1983b). The two variants have been shown to be alleles, by segregation studies (Bell and McKenzie, 1967). DNA sequencing of SS1, presented in chapter 4, shows that it encodes a histidine at amino-acid 20 (in exon II). Restriction enzyme digests of sheep genomic DNA probed with the ovine BLG cDNA, p931, and with subclones of SS1 and SS12 give band patterns consistent with a copy number of one (see figure 4.2).

Transgenic mouse lines produced by microinjection of SS1 have been described (Simons et al., 1987). Female mice transgenic for the 16.2 kb *SalI/SalI* fragment and the 10.5 kb *SalI/XbaI* fragment of SS1 secrete BLG in their milk. The 10.85 kb *SalI/XbaI* fragment of SS12 was also injected into fertilised mouse eggs, resulting in 8 G0 offspring. Four lines of mice were established from G0 mice 54, 64 and 75. Mouse 75 had integrated the injected DNA in two sites. These subsequently segregated in offspring to give lines 75L and 75H (J. P. Simons and A. J. Clark, unpublished results). Females of lines 64 and the two lines derived from mouse 75 secrete ovine BLG into their milk (M. McClenaghan, unpublished results).

SDS-polyacrylamide gel electrophoresis gives separation largely on size and does not distinguish between the two ovine BLG variants. Starch-gel electrophoresis at pH 7.6 in a phosphate buffer gave separation of ovine BLG to show two components,

named BLG-A and BLG-B (Bell and McKenzie, 1964, 1967). Conti et al. (1977) described the use of flat-bed isoelectric focusing (IEF) polyacrylamide gels to separate the two ovine BLG variants (see Materials and Methods).

Milk samples from mouse lines 45.20 (SS1) and 75L (SS12) were run on a flat-bed IEF gel, together with sheep milk samples. After focusing, the proteins were transferred to a nitrocellulose membrane and were analysed by Western blotting (see Simons et al., 1987) using an antiserum against purified ovine BLG (kindly provided by P. Gaye). Figure 5.1 (given by M. McClenaghan) shows the three BLG focusing patterns observed in sheep milk samples. The figure also shows transgenic mouse samples. From this it seems clear that SS1 (45.20) and SS12 (75L) encode different BLG types (M. McClenaghan's results).

In this chapter I present additional data showing that SS12 encodes a functional gene and show that SS12 and SS1 encode ovine BLG-A and -B, respectively. Ewes producing BLG-A or BLG-B and heterozygotes (BLG-A/B) have been used to show that the BLG phenotype is linked to a restriction fragment size difference present in SS1 and SS12.

5.1 MEASURING THE pIs OF THE TWO BLG VARIANTS

Bell et al. (1968) used tryptic digest peptides to show that ovine BLG-A contains one more tyrosine and one less histidine than ovine BLG-B. They suggested that there was also a glutamine difference between the two variants. The amino-acid sequence derived by Kolde and Braunitzer (1983a, b) indicates that there is only a histidine/tyrosine difference between the two variants.Thus BLG-A contains a tyrosine

Figure 5.1. SS1 and SS12 encode different BLGs. Shown is a Western blot of transgenic mouse milk whey, run on an isoelectric focusing gel. Shown are milk samples from transgenic mice 45.20 (labelled m1) and 75L (labelled m2). Also shown are milk samples from two sheep homozygous for BLG-B, two sheep homozygous for BLG-A and two heterozygotes. The anode is indicated (+). This figure was kindly provided by M. McClenaghan.

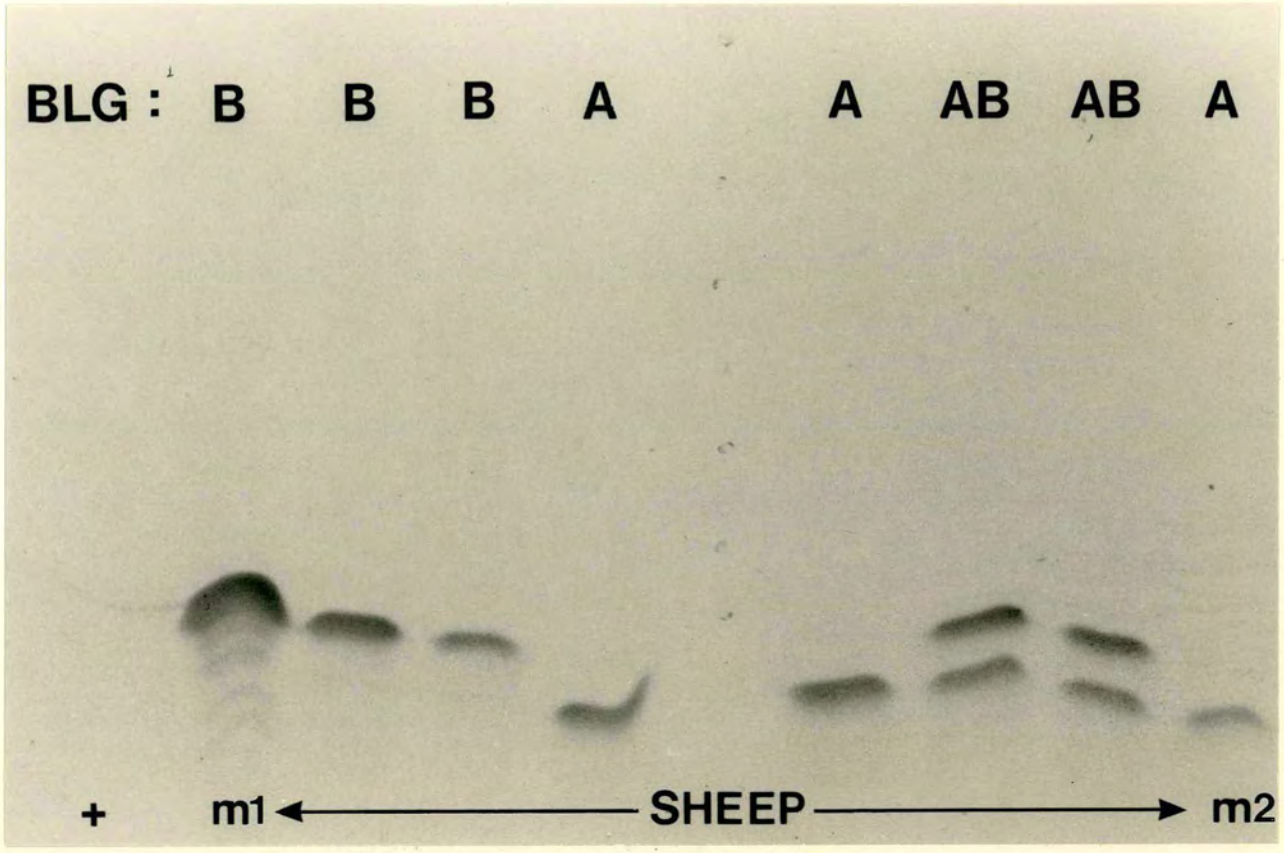
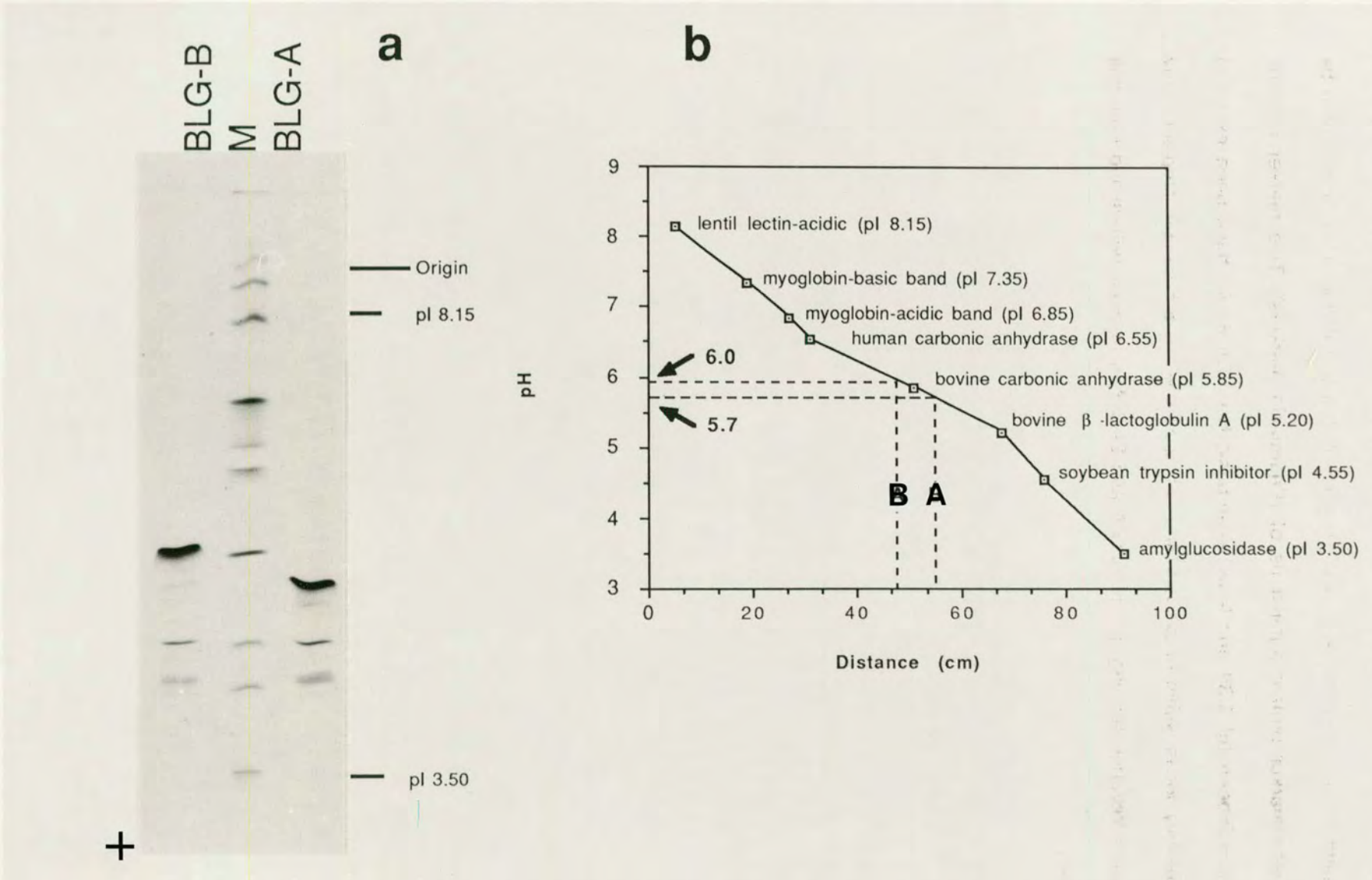

BLG : B B B A A AB AB A
+
m1
SHEEP
m2

Figure 5.2. Determination of pI values for ovine BLG-A and BLG-B. (a) is an isoelectric focusing gel showing BLG-A- and BLG-B-containing sheep milk samples. A marker track is also shown. (b) is a plot of the distance migrated by each marker protein against its pI (Pharmacia markers). Two marker proteins (trypsinogen (pI 9.30 and lentil lectin-basic band (pI 8.65)) have been excluded as they focus within or close to the cathode or within the sample strip and may show distorted running. pIs of BLG-A and BLG-B have been determined from this plot.


a
BLG-B
M
BLG-A
Origin
pI 8.15
pI 3.50
+
b
pH
9
8
7
6
5
4
3
0
20
40
60
80
100
Distance (cm)
lentil lectin-acidic (pI 8.15)
myoglobin-basic band (pI 7.35)
myoglobin-acidic band (pI 6.85)
human carbonic anhydrase (pI 6.55)
bovine carbonic anhydrase (pI 5.85)
bovine β -lactoglobulin A (pI 5.20)
soybean trypsin inhibitor (pI 4.55)
amylglucosidase (pI 3.50)
6.0
5.7
B
A

residue at amino-acid 20, whereas BLG-B contains a histidine at this residue (Kolde and Braunitzer, 1983a, b; Gaye et al., 1986). Since histidine is a basic amino-acid, whereas tyrosine is almost neutral, BLG-A should be acidic relative to BLG-B. On this basis BLG-A should focus at a higher pH than BLG-B, as is indeed the case. BLG from transgenic mouse 45.20 runs identically with the presumed BLG-B on isoelectric focusing gels. Thus, SS1 encodes BLG-B, as predicted by DNA sequencing (chapter 4). On this evidence SS12 encodes BLG-A.

To measure pIs of the two BLG variants sheep milk whey samples were run on IEF gels, together with focusing markers (from Pharmacia - see Materials and Methods) with pIs in the range 3.50 to 9.30. Plotting pI against distance moved should give an approximate linear plot. pIs of unknowns can be measured from this plot. Figure 5.2a shows an IEF with two sheep milk samples and a track showing the markers. In figure 5.2b the distance from origin is plotted against pI. This gives pI values for BLG-A and BLG-B of 5.7 and 6.0, respectively.

5.2 DNA SEQUENCING OF SS12

Kolde and Braunitzer (1983a, b) determined the amino-acid sequence of the ovine BLG variant which contains a tyrosine residue at amino-acid 20. DNA sequencing of SS1 shows that it encodes BLG-B and differs from the sequence of Kolde and Braunitzer (1983a) by a single amino-acid. To show that SS12 encodes BLG-A and to confirm that a single amino-acid difference is responsible for the difference in focusing, the exonic sequence of SS12 was obtained using the information gained from restriction mapping and sequencing of SS1.

Two subclones of SS12 were made (figure 5.3). The 4 kb *BamHI* fragment

containing exons I-III and the 2 kb *BamHI* fragment containing exon IV-VII were cloned into pUC19 (Yannisch-Perron et al., 1985). Restriction mapping with *PstI*, *PvuII*, *SmaI* and *StuI* showed only two differences between SS1 and SS12. Both are *SmaI* sites present in SS1 but absent from SS12 (figure 5.3; positions -122 and +3903 in SS, see figure 4.4 and figure 5.5 (bp 1152)). The M13 clones used are shown in figure 5.3. Arrows indicate the length of the determined sequence. The broken arrow indicates the clone constructed and sequenced by A. J. Clark. DNA sequencing showed that there are few differences between the two genes even in 5' flanking and intronic sequences.

Figure 5.4 shows the DNA sequence of the exons in SS12. The SS1 sequence is shown where it differs from that of SS12. The translated sequences show a difference at amino-acid 20. As expected, a histidine in BLG-B is encoded by a tyrosine in SS12. No other amino-acid differences are seen. The base difference which causes a histidine/tyrosine difference in SS1/SS12 is the only difference seen in the translated sequences. No silent differences are found. The 5' untranslated sequence is identical with that of SS1 but two differences are seen in the 3' untranslated sequences. One of these is a single base pair insertion/deletion, the second is a C/T difference. Figure 5.4 shows SS12 5' flanking and intron I sequences, together with SS1 sequence where it differs from that of SS12. The sequence in the region 794-1117 bp of SS12 was determined by A. J. Clark (unpublished results). Sequences 5' of 460 have not been determined for SS1. There are very few differences between the SS1 and SS12 sequence; 17 base substitutions being found in 1760 bp sequenced in SS1 and SS12 (99% homology).

Figure 5.3. Structure of the BLG gene SS12. The figure shows the detailed restriction map of the SS1 subclones pSS1BH and pSS1HX. The restriction sites used are shown. The exons are shown as open boxes; coding regions are shaded. M13 clones are shown as arrows, denoting the extent and direction of sequence. The broken arrow indicates the M13 clone constructed and sequenced by A. J. Clark.

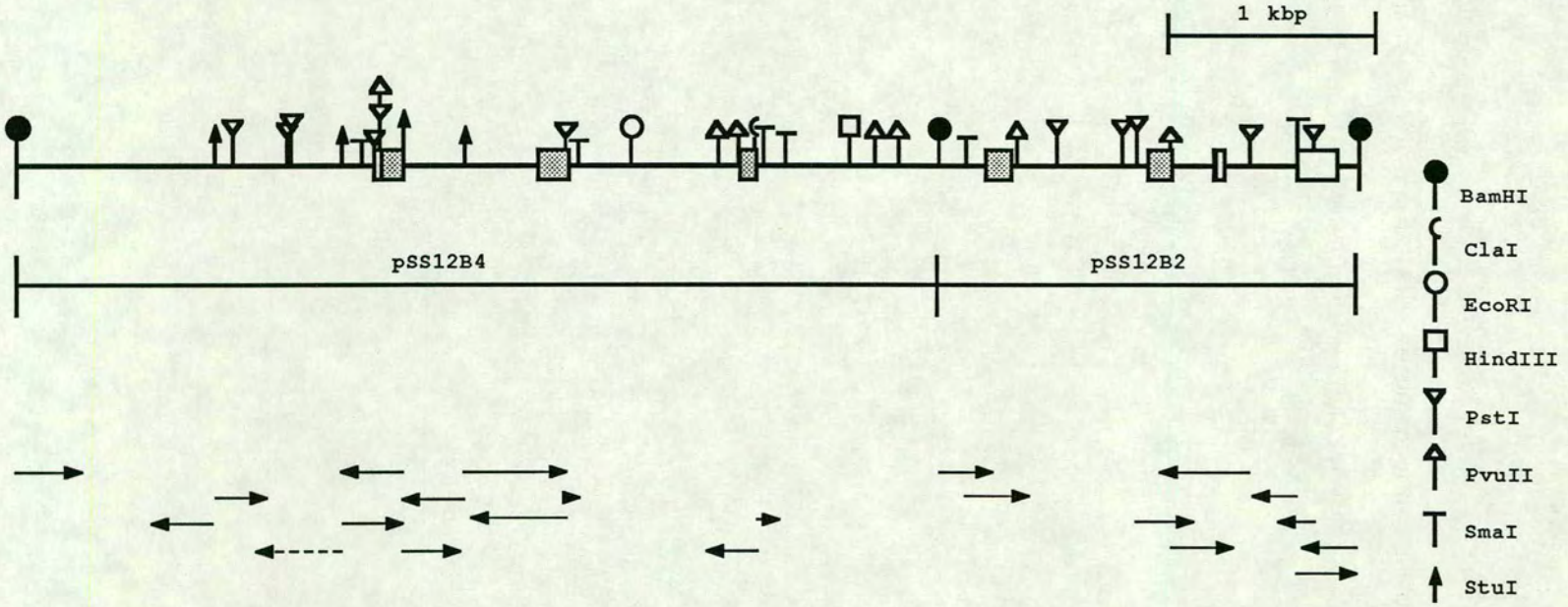

1 kbp
pSS12B4
pSS12B2
BamHI
ClaI
EcoRI
HindIII
PstI
PvuII
SmaI
StuI

Figure 5.4. Exonic sequence of BLG gene SS12. Exon sequences are shown in upper case, flanking sequences are in lower case. The predicted amino-acid sequence is shown above the DNA sequence. The numbers immediately above the amino-acid sequence are given relative to the N-terminal amino-acid of the mature BLG polypeptide. Negative numbers refer to the BLG gene signal peptide. SS1 sequence is shown immediately below the DNA sequence where it differs from SS12 sequences. (-) indicates a gap in the alignment with SS1 sequences. Putative TATA, transcriptional start site, translation start and stop and AATAAA signals are underlined (see table 4.1). The single amino-acid difference between SS12 and SS1 sequence is shown (amino-acid 20).

Exon I 136 bp

```
                                                    *
gcacgcctcctgtataaggccccaagcctgctgtctcagccctccACTCCCTGCAGAGCTCAGAAGCACGACCCCAGCTG
   t
     -18                                                        -1 +1
     MetLysCysLeuLeuLeuAlaLeuGlyLeuAlaLeuAlaCysGlyValGlnAlaIleIleValThrGlnThrMet
CAGCCATGAAGTGCCTCCTGCTTGCCCTGGGCCTGGCCCTCGCCTGTGGCGTCCAGGCCATCATCGTCACCCAGACCATG

      10
LysGlyLeuAspIleGlnLys
AAAGGCCTGGACATCCAGAAGgttcgagggt
```

Exon II 140 bp

```
                          20                                  30
          ValAlaGlyThrTrpTyrSerLeuAlaMetAlaAlaSerAspIleSerLeuLeuAspAlaGlnSerAaaP
ccctctccagGTGGCGGGGACTTGGTACTCCTTGGCTATGGCGGCCAGCGACATCTCCCTGCTGGATGCCCAGAGTGCCC
                          C
                          His
     40                             50                                  60
roLeuArgValTyrVAlGluGluLeuLysProThrProGluGlyAsnLeuGluIleLeuLeuGlnLysTr
CCCTGAGAGTGTACGTGGAGGAGCTGAAGCCCACCCCCGAGGGCAACCTGGAGATCCTGCTGCAGAAATGgtgggcgtcc
                                                                        t
```

Exon III 74 bp

```
                                    70                                  80
          pGluAsnGlyGluCysAlaGlnLysLysIleIleAlaGluLysThrLysIleProAlaValPheLysIle
tgtctttcagGGAGAACGGCGAGTGTGCTCAGAAGAAGATTATTGCAGAAAAAACCAAGATCCCTGCGGTGTTCAAGATC

AspA
GATGgtgagtccgg
```

Exon IV 111 bp

```
                     90                                  100
          laLeuAsnGluAsnLysValLeuValLeuAspThrAspTyrLysLysTyrLeuLeuPheCysMetGluAs
ctgcttccagCCTTGAATGAGAACAAAGTCCTTGTGCTGGACACCGACTACAAAAAGTACCTGCTCTTCTGCATGGAAAA
 c  g
 110                           120
nSerAlaGluProGluGLnSerLeuAlaCysGlnCysLeuV
CAGTGCTGAGCCCGAGCAAAGCCTGGCCTGCCAGTGCCTGGgtgggtgcca
```

Exon V 105 bp

```
                                 130                                 140
          alArgThrProGluValAspAsnGluAlaLeuGluLysPheAspLysAlaLeuLysAlaLeuProMetHi
tgccccatagTCAGGACCCCGGAGGTGGACAACGAGGCCCTGGAGAAATTCGACAAAGCCCTCAAGGCCCTGCCCATGCA

          150
sIleArgLeuAlaPheAsnProThrGLnLeuGluG
CATCCGGCTTGCCTTCAACCCGACCCAGCTGGAGGgtgagcaccc
                                      cg
```

Exon VI 42 bp

```
               160
          lyGlnCysHisVal***
tcccccacagGGCAGTGCCACGTCTAGGTGAGCCCCTGCCGGTGCCTCTGGGgtaagctgct
```

Exon VII 179 bp

```
ccattttcagGGCCCGGGAGCCTTGGCTCCTCTGGGGACAGATGACGTCACCACCGCCCCCCCCC-ATCAGGGGGACTA
                                                C                   C

GAAGGGACCAGGACTGCAGTCACCCTTCCTGGGACCCAGGCCCCTCCAGGCCCCTCCTGGGGCTCCTGCTCTGGGCAGCT

TCTCCTTCACCAATAAAGGCATAAACCTGTgctctcccttctgagtctttgctggacgacgggcaggggggt
```

Figure 5.5. SS12 5' flanking and intron 1 sequences. Exonic sequences are in upper case and are underlined. Exon I sequences begin at bp 1269 (see chapter 4). SS12 sequences are compared with SS1 sequences. SS1 sequences are shown immediately below the SS12 sequence weher they differ. The dotted line indicates the region of SS1 whose sequence has not been determined. The SS12 sequence in the region 794-1117 bp was determined by A. J. Clark.

5' flanking, exon I, intron I and ExonII sequences of SS12.

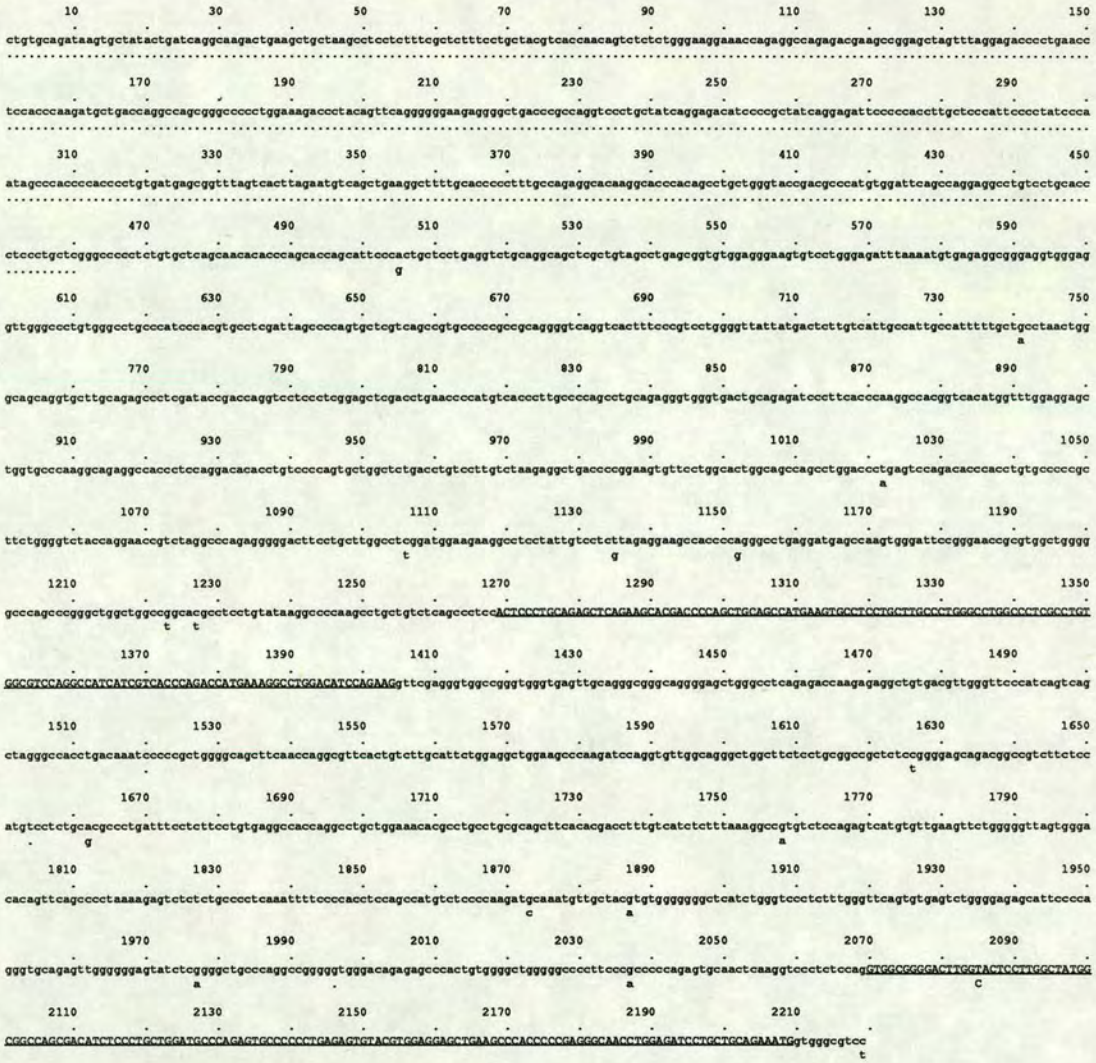
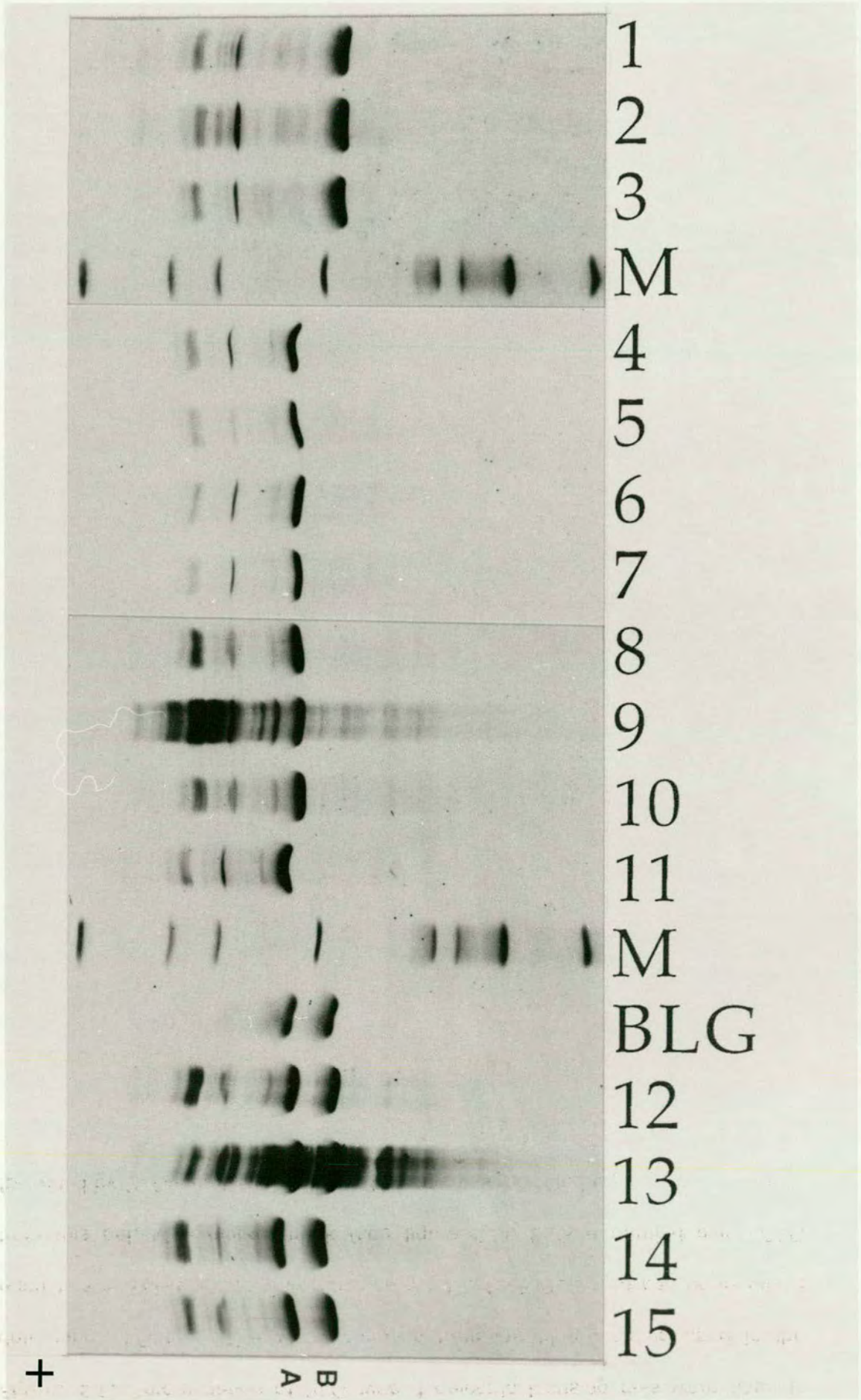


Figure 5.6. Determination of BLG type. Isoelectric focusing gels were used to determine BLG type. Milk whey samples from the fifteen sheep used in the restriction mapping analysis are numbered 1-15. Their BLG type can be identified from this gel. M = marker track (see figure 5.2). BLG = purified ovine BLG (provided by P. Gaye, INRA, France). The anode is indicated (+).


1
2
3
M
4
5
6
7
8
9
10
11
M
BLG
12
13
14
15
+
A
B

5.3 DO SS1 AND SS12 ALWAYS CORRESPOND TO THE SHEEP GENES FOR BLG-B AND BLG-A?

It has been shown that both SS1 and SS12 are functional genes, encoding the two BLG variants, BLG-B and BLG-A, respectively. To further show that SS1 and SS12 are the BLG genes present in sheep, milk and genomic DNA restriction digests were compared. Milk samples obtained from 51 sheep were analysed by IEF electrophoresis and their BLG type was determined. Some sheep producing only BLG-A, only BLG-B or BLG-A/B (heterozygotes), were used for DNA analysis. The IEF patterns obtained for the fifteen sheep are shown in figure 5.6 (and table 5.1). Sheep 1-3 appear to be homozygous for BLG-B, sheep 4-11 are homozygous for BLG-A and sheep 12-15 are heterozygotes, producing BLG-A and -B in their milk.

The genomic clone restriction maps (figure 4.1) show that a *HindIII* site present 0.2 kb 3' of the *XbaI* site in SS2 and SS12 is absent from SS1 and SS11. Genomic DNAs from the fifteen sheep in figure 5.6 were digested with *HindIII*, Southern blotted and hybridised with the 2 kb *BamHI* fragment of SS1 (which contains exons IV-VII). Figure 5.7 shows the fragments which are expected. If SS1 and SS12 are the genes present in sheep then a band of 9.7 kb should be seen in the animals producing BLG-B and a band of 4.3 kb in sheep producing BLG-A. Figure 5.8 shows the band pattern that was obtained (see also table 5.1). Sheep 1-3 should give a single band of 9.7 kb, sheep 4-11 should give a band of 4.3 kb and sheep 12-15 should give both bands. Each sheep gave either a 4.3 kb or a 9.7 kb band, or both bands. The *HindIII* pattern, however, is not concordant with the histidine/tyrosine protein types (table 5.1). The 6.7 kb band discussed in chapter 4 was seen in all of these *HindIII* digests.

HindIII digestions of the genomic DNAs were probed with the 3.4 kb *SalI/XbaI* fragment of SS12 (avoiding the extreme 3' repeat present at the 3' end in SS1) to look

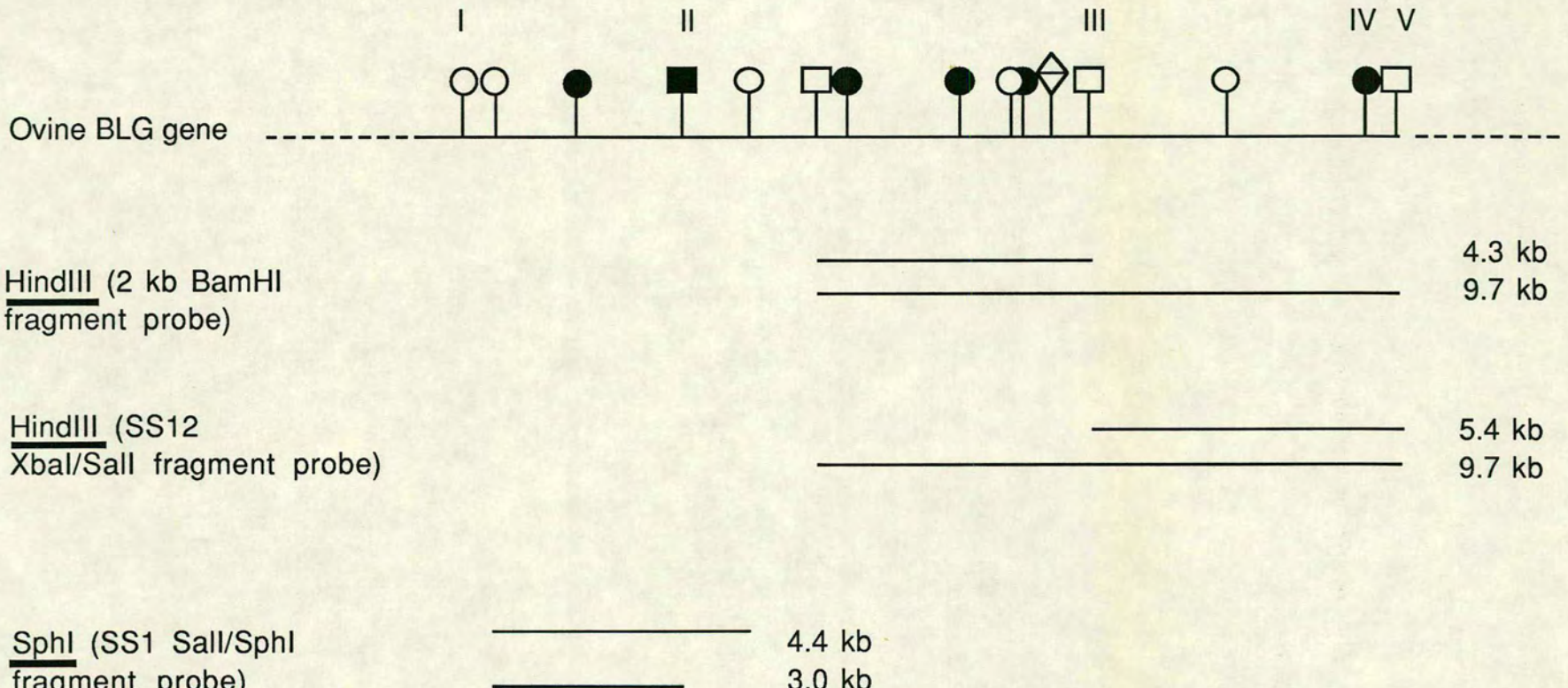

I
II
III
IV V
Ovine BLG gene
HindIII (2 kb BamHI fragment probe)
4.3 kb
9.7 kb
HindIII (SS12 XbaI/SalI fragment probe)
5.4 kb
9.7 kb
SphI (SS1 SalI/SphI fragment probe)
4.4 kb
3.0 kb

Figure 5.7. Ovine BLG restriction map and expected restriction digest fragments. The restriction map of the ovine BLG gene is shown. All restrction enzyme sites are indicated.The known polymorphic restriction sites are numbered I-V. Also shown are the polymorphic restriction fragments (and their sizes) expected with *HindIII* and *SphI* digestions.

Figure 5.8. *HindIII* digestion of sheep genomic DNAs. Genomic DNAs from the fifteen sheep were digested with *HindIII*, run on a 0.8% agarose gel, Southern blotted, and probed with the 2 kb *BamHI* fragment which contains BLG gene exons IV-VII. The two polymorphic bands are arrowed (4.3 kb and 9.7 kb). The 6.7 kb band (*) is discussed in chapter 4 (figure 4.2). Copy control *HindIII* digests of SS1 and SS12 phages are also shown.

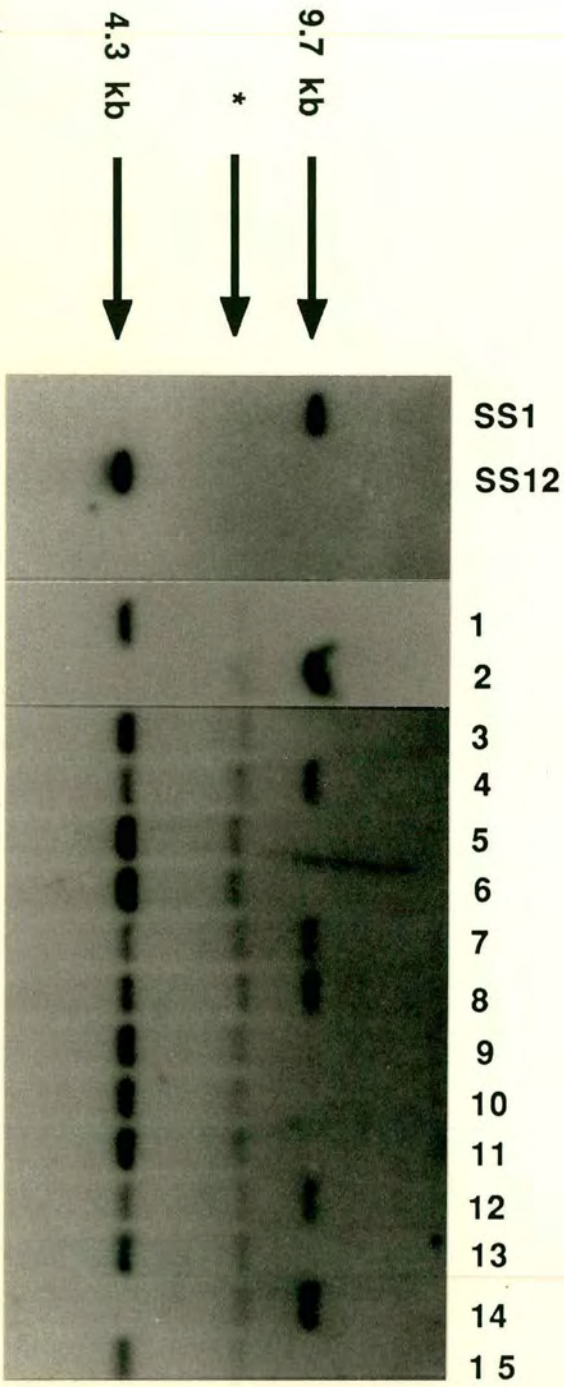

4.3 kb
*
9.7 kb
SS1
SS12
1
2
3
4
5
6
7
8
9
10
11
12
13
14
1 5

Figure 5.9. *HindIII* digestion of sheep genomic DNAs. Genomic DNAs from the fifteen sheep were digested with *HindIII*, run on a 0.8% agarose gel, Southern blotted, and probed with the 3.4 kb SS12 *SalI/XbaI* fragment. Copy control *HindIII* digests of SS1 and SS12 phages are also shown. The numbers are lambda marker sizes, in kb.

SS1
SS12
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
21.3
5.15/4.97
4.27
3.54

Figure 5.10. *SphI/EcoRI* digestion of sheep genomic DNAs. Genomic DNAs from the fifteen sheep were digested with *SphI/EcoRI*, run on a 0.8% agarose gel, Southern blotted, and probed with the 4.0 kb SS12 *SalI/SphI* fragment. Some digests are shown twice. Lambda marker sizes are shown, in kb. The 3.0 and 4.4 kb bands are asterisked.

2
1
3
4
5
6
7
8
9
10
11
12
13
14
15
4
12
13
14
21.3
5.15/4.97
4.27
3.54
2.03/1.90
1.58
1.38

for the presence or absence of the *HindIII* fragment, which is present in SS1 and SS11 but apparently absent from SS2. Figure 5.7 shows that if the 3' site is present fragments of 5.4 kb and/or 9.7 kb should be found. A band greater than 5.4 kb in size is expected for a SS12-like gene. If a SS1/SS11-like gene is present a 9.7 kb *HindIII* fragment should be seen, whereas a fragment of greater than 6.25 kb should be seen in a SS2/SS12-like gene. Figure 5.9 shows the fragments which were seen. Those DNAs which give a 9.7 kb *HindIII* fragment with the 2 kb *BamHI* fragment as a probe, also show this fragment, as expected. However, a 5.4 kb band was seen in the remainder of the digests. This means that all the sheep (not known for sheep no. 6) contain BLG genes which have the *HindIII* site present at the very 3' end of SS1. Thus, the absence of this *HindIII* site from SS2 appears to be unusual. It is, however, possible that the difference between SS2 and other sheep BLG genes is due to a cloning artifact. Exhaustion of the DNA used to make the sheep genomic library prevents analysis to determine whether the SS2 pattern is a cloning artifact.

A third restriction enzyme site polymorphism has been examined. A *SphI* site is present at -42 in SS1 but is absent from SS12. DNA sequencing shows that a 'transitional' C/T difference accounts for this (the first six bases in figure 5.4; position 1227 in figure 5.5). No *SphI* sites are found downstream of the mapped SS1 site in either SS1 or SS12, for at least 12 kb, and upstream for at least 4 kb. There is no SphI site in SS11 suggesting that there is no *SphI* site 3' of the one in SS1 for more than 17 kb. This suggested that it would be difficult to assay *SphI* differences due to the large fragment sizes likely to be obtained from both SS1- and SS12-like genes. Therefore, *SphI/EcoRI* genomic DNA digests were probed with SS1 *SalI/SphI* fragment. The absence of a *SphI* site will yield an *EcoRI* fragment of 4.4 kb. If a *SphI* site is present 1.4 and 3.0 kb bands will be obtained (only the 3.0 kb band will be seen with the *SalI/SphI* probe). Figure 5.10 shows the band pattern obtained. All fifteen DNAs

gave fragment sizes predicted from SS1 and SS12 restriction maps. Thus, the *SphI* polymorphism appears to be linked to BLG variants. Two other bands (about 1.0 and 2.8 kb in size) were seen in these digests. It is not clear which of these is the fragment expected from the presence of *EcoRI* or *SphI* sites in sequences flanking the 5' ends of the two cloned phage types. The presence of both bands in all the digests (both bands are not clearly seen in all samples presented here, but are seen in longer exposures) suggests that one of these fragments arises from hybridisation to the "related" sequences described in chapter 4. In addition, the 0.4 kb *EcoRI* fragment expected from SS1-like BLG genes is not seen in any of these digests. It is possible, however, that the gel conditions did not resolve this fragment. Preliminary results using *EcoRI* digests probed with the 0.4 kb *EcoRI* fragment indicate that it is indeed a polymorphic site (also see discussion).

The *SphI* polymorphism (this site is present about 650 bp upstream of the histidine/tyrosine codon, in SS1) appears to show linkage with the protein polymorphism types in this population. The *HindIII* pattern is not linked to the protein polymorphism (the *HindIII* site present in SS12 is situated about 6.5 kb downstream of the polymorphic codon, the second polymorphic *HindIII* site is a further 5.4 kb downstream).

Table 5.1 The correlation between BLG protein alleles and haplotypes.

Sample	Variant	*SphI*	*HindIII(I)*	*HindIII(II)*
SS1	B	+	-	+
SS12	A	-	+	-
1	B/B	+/+	+/+	+/+
2	B/B	+/+	-/-	+/+
3	B/B	+/+	+/+	+/+
4	A/A	-/-	+/-	+/+
5	A/A	-/-	+/+	+/+
6	A/A	-/-	+/+	NK
7	A/A	-/-	+/-	+/+
8	A/A	-/-	+/-	+/+
9	A/A	-/-	+/+	+/+
10	A/A	-/-	+/+	+/+
11	A/A	-/-	+/+	+/+
12	A/B	-/+	+/-	+/+
13	A/B	-/+	+/+	+/+
14	A/B	-/+	-/-	+/+
15	A/B	-/+	+/+	+/+

The table shows the type of BLG secreted by each sheep, 1-15, and summarises the data shown in figures 5.7, 5.8 and 5.9. The presence of each polymorphic restriction site is indicated by (+), its absence by (-). NK = Not Known.

Table 5.2 BLG Gene haplotypes.

Haplotype	SphI(II)	HindIII(III)	HindIII(V)	Phenotype
I	+	+	+	B
II	+	-	+	B
III	-	+	+	A
IV	-	-	+	A
V	-	+	-	A

The five haplotypes described in this chapter, are shown. (+) indicates the presence and (-) indicates the absence of a restriction site. The phenotypes which have been observed for each haplotype are shown.

5.4 DISCUSSION

The transgenic mice data clearly indicate that both SS1 and SS12 are functional genes which direct expression of the ovine BLG gene in the mammary gland, to produce ovine BLG in mouse milk. SDS-polyacrylamide gel electrophoresis shows that BLG is present in mouse milks from both transgenics. The BLGs can be distinguished by isoelectric focusing. Thus SS1 and SS12 encode similar, but distinguishable, proteins.

To determine pIs of the two proteins IEF gels were run, with pI markers. A plot of pI versus distance moved from the origin, gives pIs of 5.7 and 6.0 for BLG-A and BLG-B, respectively.

It has been shown previously (Conti et al., 1977) that the two variants of BLG present in sheep milk can be separated on IEF gels. It has also been shown that the two variants are alleles (Bell and McKenzie, 1967). Peptide analysis (Bell et al., 1968) and amino-acid sequencing (Kolde and Braunitzer, 1983a, b) have suggested that the difference between BLG-A and BLG-B involves a tyrosine residue at amino-acid 20 (in BLG-A), although it was not clear which amino-acid replaces it in BLG-B. Gaye et al. (1986) obtained a cDNA clone (p931), encoding a histidine residue at amino-acid 20 (consistent with encoding BLG-B). However, their cDNA is incomplete, only stretching to amino-acid 10. Primer extension enabled determination of the remainder of the mRNA sequence but mammary RNA from a different sheep, which encoded a tyrosine residue at amino-acid 20 (the variant whose amino-acid sequence was determined by Kolde and Braunitzer (1983a)), was used. The possibility of another difference between BLG-A and BLG-B, in amino-acids 1-10 could not be ruled out. Sequencing of SS1 and SS12 shows that the histidine/tyrosine difference is the only amino-acid difference between BLG variants A and B. SS1 and SS12 show very

few differences, most of which are base substitutions. A homology of 99% is found in the 5' flanking and intron sequences.

A great deal of work has been carried out on genetic polymorphisms, first on protein polymorphisms (including BLG polymorphisms - see introduction to this chapter) describing electrophoretic differences, and after the discovery of restriction enzymes much work has involved study of DNA polymorphisms. Restriction site differences define polymorphisms which are most often silent at the phenotypic level by their presence in flanking DNA, introns or at silent positions in codons. Hence, much more heterogeneity can be present at the DNA level without affecting protein structure. Much of the initial work utilising restriction enzymes was done on the human β-globin genes (Lawn et al., 1978; Kan and Dozy, 1978; Jeffreys, 1979). Jeffreys (1979) used eight different restriction enzymes and defined three polymorphisms (by the presence or absence of restriction enzyme cleavage sites) over a region of 41 kb. Subsequent work has detected 17 common polymorphisms at the β-globin locus. Almost all of the discovered polymorphisms lie in flanking and intronic DNA (reviewed by Orkin and Kazazian, 1984). DNA polymorphisms at the β-globin locus have also been mapped in different population groups. Some polymorphisms are common to all populations, suggesting their ancient origin, whilst others are more common in certain populations (see Orkin and Kazazian, 1984). Furthermore, Antonarkis et al. (1982) suggested that polymorphic sites are not randomly associated with each other but that certain combinations of the polymorphic restriction enzyme sites are found, generally a far fewer number than are possible. The arrangements of restriction sites found are termed haplotypes. Some haplotypes are more common in certain populations than in others. Much of the work on the β-globin locus described in this paragraph generally agrees with findings for other genes (for example, HLA class I genes - Ehrlich et al., 1983; type II procollagen gene - Eng and Strom, 1985). In

particular, much interest has centred around locating DNA polymorphisms showing linkage to human genetic diseases (Sickle cell anemia - Kan and Dozy, 1978; Antithrombin III deficiency - Prochownik et al., 1983; Growth hormone deficiency type II - Phillips et al., 1981; Phenylketonuria - Woo et al., 1984; various thalassemias and many other conditions - see Caskey (1987) for review) which could enable prenatal diagnoses to be attempted

Jeffreys (1979) used the restriction enzyme data to suggest that about 1 bp in 100 is polymorphic in the β-globin locus. Subsequent work with the β-globin locus appears to agree with these findings (see Orkin and Kazazian, 1984). Furthermore, Slightom et al. (1980) sequenced both chromosomal copies of the human δ-globin gene. The two sequences differed at 18 positions (15 base substitutions, 3 small insertion/deletions) over a sequence of just under 1500 bp (this includes two exons (299 bp) which show no differences in sequence), agreeing reasonably well with Jeffreys' (1979) suggestion.

Less work has been done in species other than man and in any case most other species analysed have been inbred for research (for example *Drosophila* and rodents) or for agriculture (for example cattle and sheep). Kreitman (1983) sequenced the alcohol dehydrogenase gene of different *Drosophila* strains; sequencing genes which encode electrophoretically non-separable proteins. He found 43 polymorphisms in 1721 bp (1.8% polymorphism) in eleven sequenced alcohol dehydrogenase genes from 5 natural populations; although silent changes in codons and intron 2 and 3 sequences showed about 6% polymorphism, whilst flanking, intron 1 and untranslated sequences showed much lower polymorphism. In ruminants some restriction enzyme site polymorphisms have been mapped in sheep (Vaiman et al., 1986) and cows (for example, Beckmann et al., 1986). Beckmann et al. (1986) found four polymorphic sites in a total of 51 restriction fragments mapped (1.3 polymorphic sites per 100

bp, using the calculation made by Jeffreys (1979)), similar to results found for the human β-globin locus.

51 milk samples from 4 sheep breeds were analysed by IEF to determine their BLG protein types. No attempt, apart from the use of different breeds, was made to eliminate founder animal effects in the analysis. 3 BLG-B homozygous, 23 heterozygous and 25 BLG-A homozygous animals were identified. In more detailed studies many workers have shown that allele distribution varies in different breeds, BLG-A being more common. King (1969) compared different sheep breeds and showed that allele distribution varies from 0.50 BLG-A (Wiltshire Horn) to 0.91 BLG-A (Dorset Horn). DNA samples from various breeds, but including the three BLG-B sheep were used for this analysis. The DNA polymorphism study was carried out to determine linkage between the presence or absence of restriction enzyme sites and BLG-type, using restriction enzymes whose polymorphism was suggested by differences between SS1/SS11 and SS2/SS12. A larger number of animals would need to be examined in order to determine frequencies of polymorphic sites. Nevertheless, this limited analysis shows many of the different haplotypes which are likely to be found. Furthermore, DNA sequence comparison of SS1 and SS12 shows that about 1 bp in 100 is polymorphic, in agreement with previous estimates described above (see below for discussion of the relationship between SS1 and SS12 DNA sequences).

DNAs from fifteen sheep homozygous for BLG-A, BLG-B or heterozygous (BLG-A/B) were used for the DNA studies. Some of the restriction enzyme site differences between SS1/SS11 and SS2/SS12, were used (see figures 4.1 and 5.7). If SS1 and SS12 encode the two ovine BLG alleles restriction digestion with the appropriate enzymes may show linkage of restriction enzyme site differences to the protein polymorphism. *HindIII* digestion, however, fails to show linkage with the A/B variation. The presence or absence of a *SphI* site, on the other hand, appears to be

entirely linked with the A/B variation (see table 5.1). Probing for the presence of the 3'-most *HindIII* site of SS1 shows that all fifteen sheep BLG genes contain this site. Nevertheless, it is absent from SS2 and probably maps yet another polymorphic restriction site in the BLG gene (although cloning artifacts cannot be ruled out). Other restriction site differences between SS1 and SS2/SS12 include a *BamHI* (IV) site present at the 3' end of SS1 and two *SmaI* sites present in SS1 but absent from SS12. Preliminary evidence suggests that the *EcoRI* (I) (figure 5.7) is also polymorphic. The data suggests that there is no linkage of this polymorphism with BLG phenotype or with the other three polymorphic sites analysed.

The above analysis suggests that the BLG locus is quite polymorphic. Five out of fifteen mapped sites, over a region of 17.5 kb, are polymorphic (using five restriction enzymes - figure 5.7). Although the sample size is too small to demonstrate which haplotypes are more, or less, common the presence of some haplotypes in the population has been demonstrated. Using the Jeffreys' (1979) calculation, assuming a single base pair change at polymorphic sites, the BLG gene shows (5/(6 x 15) =) 5.6% polymorphism. This is much greater than is indicated by sequence comparisons of SS1 and SS12, but is not unusual, considering the results of Kreitman (1983) who showed 6% polymorphism in intron 2 and 3 sequences and in silent codon changes in the *Drosophila* alcohol dehydrogenase gene.

Comparison of SS1 and SS12 5', exon I, intron 1 and exon II sequences (figure 5.4) shows 17 base substitutions in 1760 bp of sequence. The number of base substitutions is about 1 per 100 bp, consistent with the findings of other workers, described above. 71% (12 out of 17) of these substitutions destroy a CG dinucleotide in SS1 or SS12. Methylated C is highly mutable to give T (Bird, 1986). The 1% polymorphism observed in these sequences, therefore, appears to be low considering that mutation of CGs to TG or CA appears to be more frequent than other mutations and

suggests that the BLG gene may be under-methylated. This low polymorphism, particularly when compared with the restriction site polymorphisms suggests that the distal 5' flanking and transcription unit sequences may be more highly conserved, relative to further upstream and downstream regions, which exhibit greater polymorphism (restriction enzyme sites I, III, IV and V).

It is clear from these data that the two ovine BLG alleles are encoded by a greater number of genes, shown by the larger number of haplotypes (table 5.2). At least five haplotypes have been shown to exist in this study. The distribution of these haplotypes would require more extensive work, describing related and unrelated animals (as described by Orkin and Kazazian (1984) for the β-globin genes; and by many others).

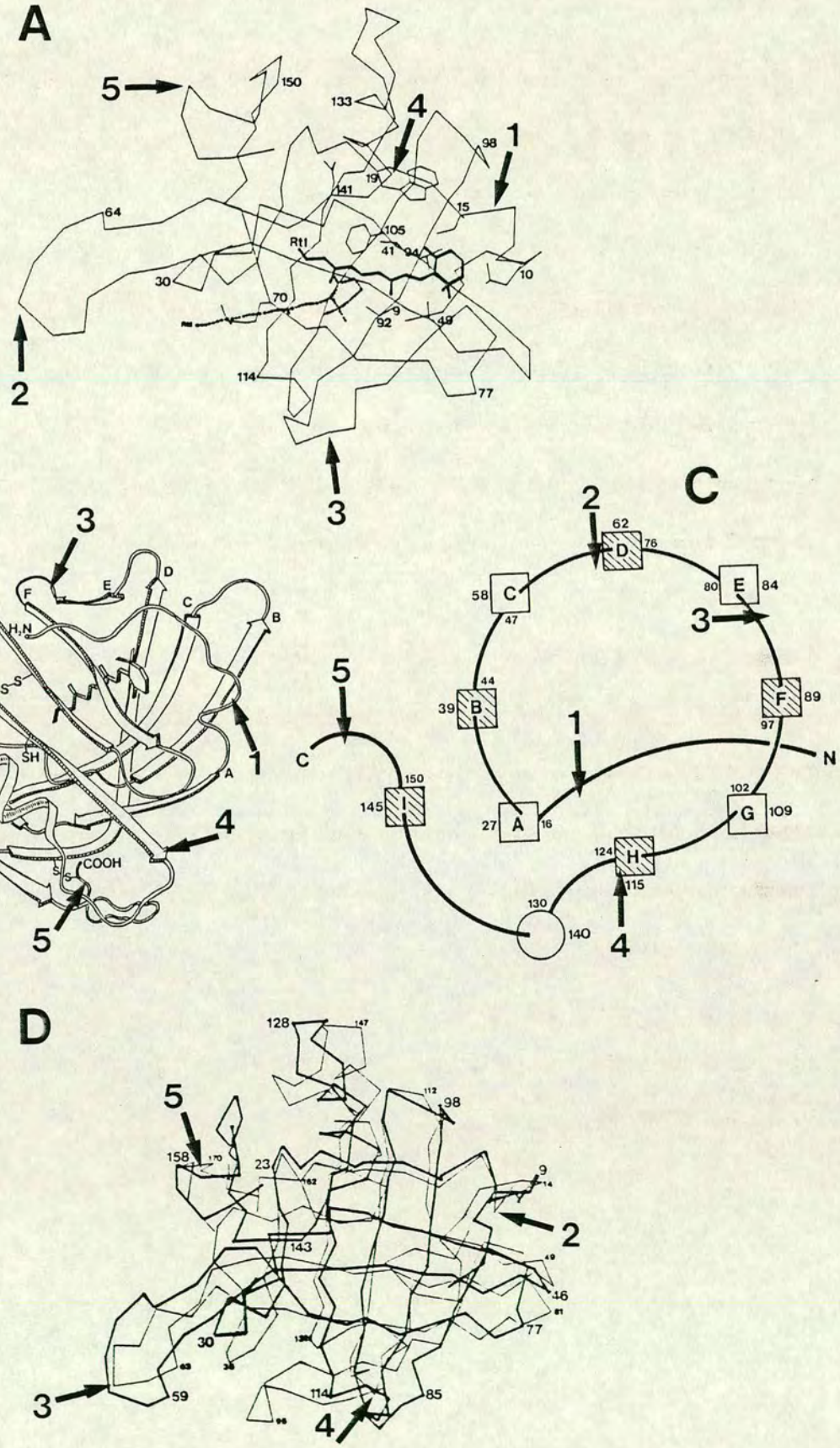
Chapter 6. COMPARISON OF β-LACTOGLOBULIN WITH OTHER SECRETORY PROTEINS

6.1 INTRODUCTION

Large amounts of β-lactoglobulin are secreted into sheep milk. However, its function is not clear and its absence from the milks of some mammals suggests that either its function is carried out by another protein(s) present in the milks of other mammals, or that its function is carried out by a β-lactoglobulin-like protein, perhaps during pregnancy. Alternatively, its role is not important or has been lost during evolution and it now remains in some mammals, purely as a source of protein for the young. A number of different functions for β-lactoglobulin have been suggested. These include transport of immunoglobulins to the young (Butler, 1974) and the transport of vitamin A or similar molecules to the young (Futterman and Heller, 1972; Fugate and Song, 1980; Papiz et al., 1986). Its stability at low pH may allow much of it to pass intact through the stomach to the intestine (Papiz et al., 1986). This property could be important for any of the above possible functions. β-lactoglobulin has been shown to be homologous to at least fourteen proteins found in insects, amphibians, birds and mammals (Sawyer, 1987; Ali and Clark, 1988). In this chapter I show that these proteins are members of a highly diverged family of proteins at least three of which share similar three dimensional structures (Papiz et al., 1986; Huber et al., 1987; Holden et al., 1987) and show that they share similar gene structures (where known). These similarities indicate possible functions for β-lactoglobulin.

Figure 6.1. The three-dimensional structure of β-lactoglobulin. The figure shows three representations of the BLG structure. (a) shows a line representation of the three-dimensional structure of bovine BLG. This figure shows the amino-acid α-carbon backbone structure. Also shown are the exon/intron junctions, intron numbers are shown (see also figure 6.3). Surface loops are more clearly seen in this figure than in the schematic representations in (b) and (c). (b) shows the β-strands and the α-helix making up the BLG structure. Arrows indicate β-strands (lettered A-I). The α-helix is shown as a coil. Lines indicate N- and C-terminal coils and connecting loops. Cysteines are indicated and the disulphide bridges are shown. (c) shows a cartoon of the β-barrel structure formed by the BLG β-strands. Boxes represent β-strands, the circle represents the α-helix. The antiparallel nature of the β-strands is indicated by cross-hatched and clear boxes. Lines indicate N- and C-terminal coils and connecting connecting loops. In (d), the three-dimensional structure of RBP is shown (heavy line), superimposed onto the BLG structure (faint line). Rat RBP gene introns are shown for comparison with (a).

These figures have been modified from those kindly provided by Dr. L. Sawyer (Dept. of Biochemistry, Univ. of Edinburgh), by the addition of exon/intron information.


A
5
150
133
4
98
1
19
141
15
64
105
Rtl
41
94
10
30
70
9
92
49
114
77
2
3
B
3
G
F
E
D
C
B
H
H,N
A
S-S
SH
I
A
1
4
COOH
2
5
C
2
62
D
76
58
C
47
80
E
84
3
5
44
39
B
F
89
97
1
N
C
150
145
I
27
A
16
102
G
109
124
H
115
130
140
4
D
128
147
5
112
98
158
170
23
162
9
14
2
143
49
16
81
77
30
63
59
114
85
3
4

6.2 COMPARISON OF THE GENE TO THE PROTEIN

Soon after the first discovery of introns Gilbert (1978) proposed that split genes may allow more rapid evolution, by exon shuffling; leading to the production of new proteins. He suggested that exons may encode "functions" so that assembly of a new assortment of exons would have a greater chance of encoding a functional novel protein. Blake (1978) modified this proposal by suggesting that exons in fact probably encode domains or supersecondary structures, rather than functions. Supersecondary structures are local organisations of secondary structure motifs, α-helices, β-strands, etc., whilst domains are more extensive structures which appear to form independent motifs within proteins (for review see Blake, 1985). Examples which have been extensively studied include immunoglobulin, ovomucoid, β-crystallin, globin and lysozyme genes (see Blake (1985) for review). If exons encode structural units then intron/exon junctions should map to regions between structural motifs. Furthermore, Craik et al. (1982) suggested that exon/intron junctions map to protein surface and that "intron sliding" may lead to changes in protein size at these regions (Craik et al., 1983).

Papiz et al. (1986) determined the 3-dimensional folding of bovine BLG. They showed that two slabs of antiparallel β-sheet form an unusual "β-barrel" structure. Figure 6.1 shows three representations of the structure (modified from Papiz et al., 1986). Figure 6.1a is a drawing of the structure which outlines the positions of the amino-acid α-carbon backbone. Figure 6.1b shows a drawing of the β-barrel structure to point out the β-strands and α-helix and to show how the barrel forms. Figure 6.1c is a schematic of the β-strands, their antiparallel nature highlighted. Ovine and bovine BLGs differ in 5 out of 162 amino-acids so their 3-D structures will be very similar. Furthermore, bovine BLG gene has been found to

have the same structure as the ovine gene (A. G. Mackinlay, personal communication). Eight antiparallel β-strands (A-H) comprise the β-barrel core. A single α-helix is present and a β-strand (I) is involved in dimer formation by interacting with the I strand in the other subunit.

Positioning the ovine BLG gene intron/exon junctions on the protein structure (figure 6.1) shows that each intron/exon junction maps to, or at, the boundaries of connecting loops separating structural units. Thus exon I encodes the signal peptide (not shown in figure 6.1) and the N-terminal 14 amino-acids of the mature polypeptide. Exon II (β-strands A-C), exon III (β-strands D and E) and exon IV (β-strands F-H) encode the β-strands which comprise the β-barrel core. Exon V encodes the α-helix and β-strand I; translation terminates in exon VI.

6.3 β-LACTOGLOBULIN IS HOMOLOGOUS TO SERUM RETINOL-BINDING PROTEIN

Pervaiz and Brew (1985) and Godovac-Zimmerman et al. (1985) first identified the homology between BLG and human serum retinol-binding protein (RBP). RBP is synthesised in the liver and secreted into the blood. It transports retinol in the blood (Laurent et al., 1985; and references therein). The homology appears to be about 20% (see figure 6.7). Figure 6.2 shows a UWGCG program DOTPLOT (Devereux et al., 1984) comparison of ovine BLG and human RBP. Low sequence homology extends over the greater length of the two proteins, but there are strong local regions of homology. Four regions of sequence similarity are present across the length of the two proteins (also see figure 6.7). DOTPLOT comparison of

Figure 6.2. DOTPLOT analysis of BLG and RBP amino-acid and cDNA sequences. (a) shows a dotplot of the amino-acid sequences of BLG and RBP (window = 20, stringency = 8.0). Four regions of sequence similarity were found across the lengths of the two proteins. An additional, relatively long region of similarity was seen (at BLG amino-acids 45-70, RBP amino-acids 125-150) (see section 6.8). (b) compares the cDNA sequences of BLG and RBP (window = 24, stringency = 14). The dotplot also shows the positions of exon/intron junctions. Little sequence similarity is seen. Nevertheless, across the lengths of the protein coding regions (boxed), some short regions of sequence similarity are seen. These appear to correspond to the regions observed in (a).

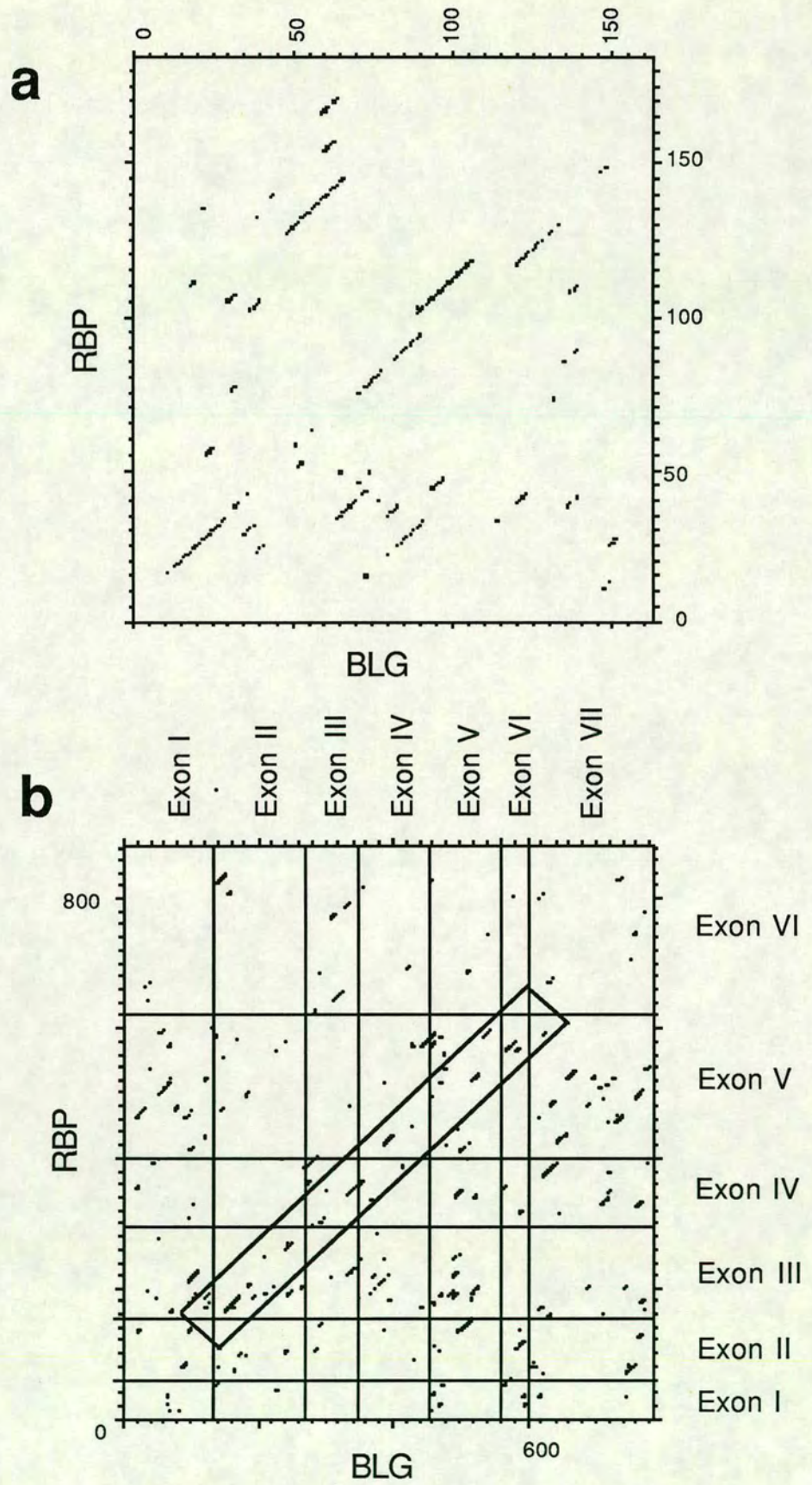
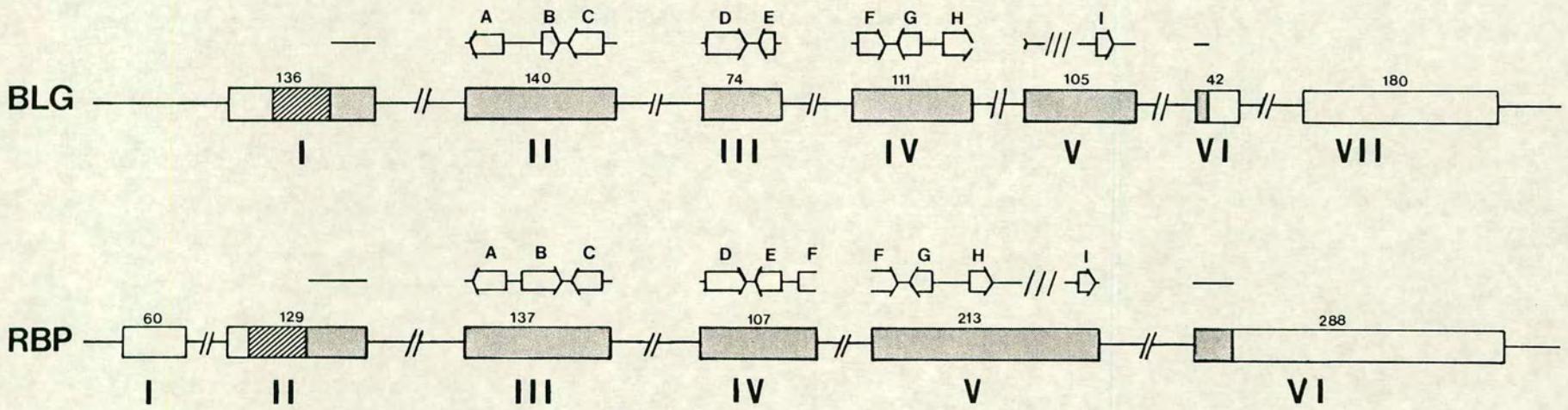

a
0
50
100
150
150
100
50
0
RBP
BLG
b
Exon I
Exon II
Exon III
Exon IV
Exon V
Exon VI
Exon VII
800
0
600
RBP
BLG
Exon VI
Exon V
Exon IV
Exon III
Exon II
Exon I

Figure 6.3. Comparison of the three-dimensional elements of BLG and RBP with their gene structures. Exons are shown as open boxes, protein coding regions are shaded (the signal peptides regions are hatched). Exon sizes (bp) are shown above the exons. Corresponding three-dimensional elements are indicated above each exon (after Laurent et al., 1985). β-strands are lettered A to I and their antiparallel nature is indicated by arrows. N-terminal and C-terminal coils and connecting loops are shown as horizontal lines and the three-turn α-helix is represented by three slashes. BLG, ovine β-lactoglobulin; RBP, rat retinol binding protein (Laurent et al., 1985).


BLG
A
B
C
D
E
F
G
H
I
136
140
74
111
105
42
180
I
II
III
IV
V
VI
VII
RBP
A
B
C
D
E
F
F
G
H
I
60
129
137
107
213
288
I
II
III
IV
V
VI

the cDNA sequences of ovine BLG and human RBP is shown in figure 6.2b. Also shown are exon/intron junctions. Short regions of homology are seen, spread over the protein-coding sequences (boxed region). Nevertheless, the homology at the DNA level is low.

Papiz et al. (1986) compared the 3D structures of ovine BLG and human RBP. They found that the two proteins have similar structures (see figure 6.1d). Alignment of the BLG and RBP amino-acid backbones showed that much of the two structures could be superimposed. 129 out of the 162 amino-acid residues could be closely aligned. The major differences were seen at external loop regions. In both proteins eight β-strands (A-H) form a β-barrel structure.

Because BLG and RBP are clearly homologous their gene structures were compared (rat RBP gene - Laurent et al., 1985). Figure 6.3 shows a schematic alignment of the exons of the two genes. Since the proteins share similar 3D structures these were compared in the context of the gene structures. The comparison shows that the two genes have similar organisations, with analogous exons encoding similar structural motifs. Exon I of the BLG gene and exon II of the RBP gene both encode the signal peptide and a stretch of N-terminal amino-acids. Exon II of the BLG gene and exon III of the RBP gene are very similar in size, differing by three nucleotides. They both encode the first three β-strands of the β-barrel. Both exons end in the reverse turn between β-strands C and D. Exon III of the BLG gene and exon IV of the RBP gene encode β-strands D and E. Exon III of BLG ends in a reverse turn between β-strands E and F. Exon IV of RBP, on the other hand, is 33 nucleotides longer and encodes part of an extended β-strand F. The exon terminates just prior to a β-bend.

The BLG gene exon IV encodes β-strands F, G and H and exon V encodes the α-helix and β-strand I. However, the RBP gene exon V encodes all five of these units. This, together with a difference in size of only three nucleotides between the RBP gene

exon V and the combined size of the BLG gene exons IV and V, indicates that an intron insertion or deletion may have occurred.

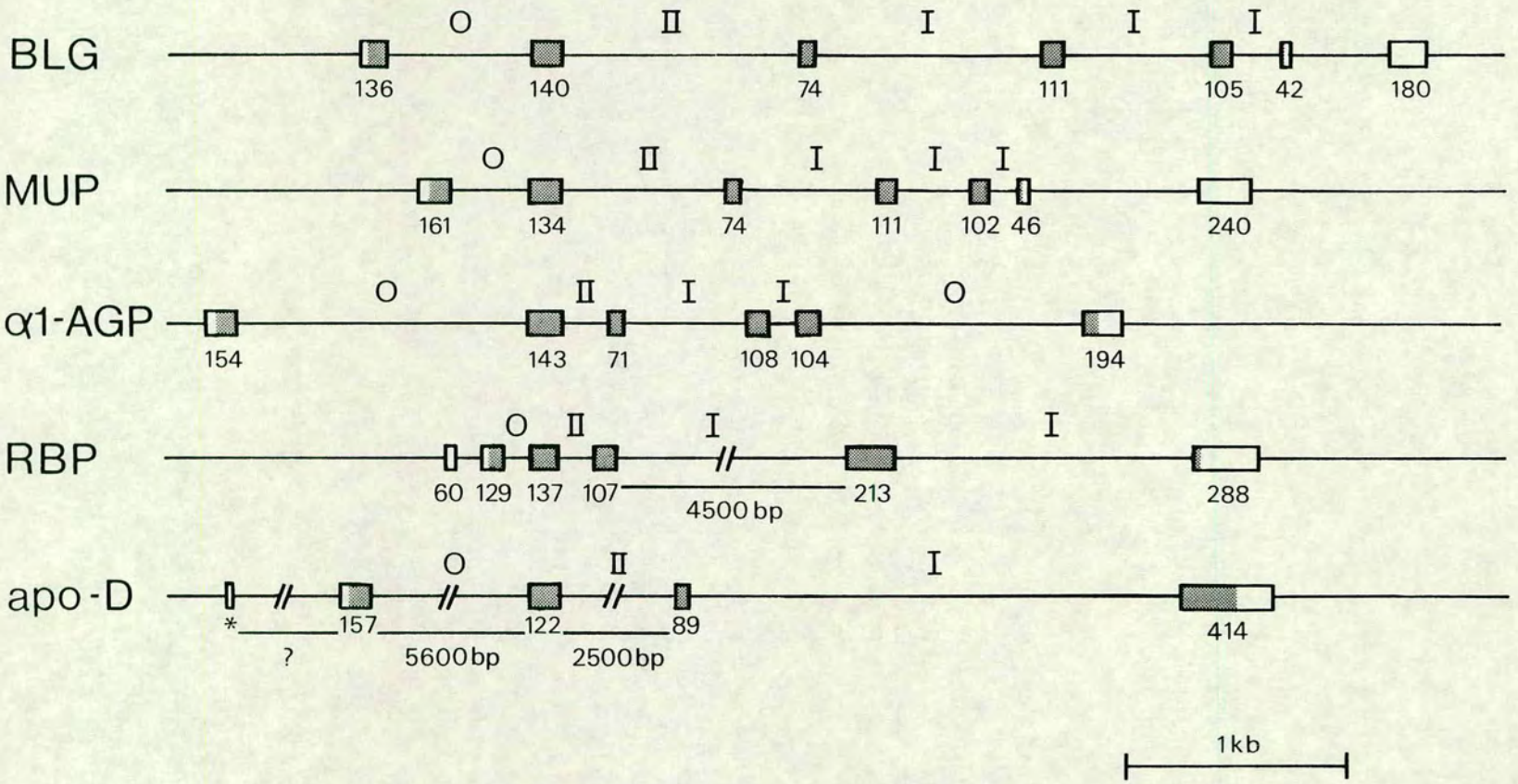
Exon VI of both the BLG and RBP genes contains a short disordered segment of amino-acids and the termination codon. In both cases, exon VI contains a cysteine residue which forms a disulphide bridge with a cysteine residue present in β-strand D.

It can be seen from this analysis that the BLG and RBP genes have similar 3D structures and gene organisations, similar secondary structure units being present within analogous exons. Also compare figures 6.1a and d. The two major differences in the gene structures have been described above. In the next section the gene structures of proteins which show amino-acid sequence similarity with BLG and RBP are also compared. The differences between the two gene structures will be discussed in the light of information on those structures.

6.4 OTHER GENES IN THE FAMILY ALSO HAVE SIMILAR GENE STRUCTURES

Limited amino-acid sequence homology between BLG and the rodent urinary globulin (Unterman et al., 1981), between BLG, RBP, rodent urinary globulins and apolipoprotein-D (Drayna et al., 1986) and between BLG, RBP, rodent urinary proteins and α1-acid glycoprotein (J. O. Bishop, personal communication) have been described. The structures of the genes encoding these proteins were compared to the BLG and RBP genes (figure 6.4). The mouse major urinary protein (MUP) gene (Clark et al., 1984) was used as a representative rodent urinary protein. It has a very similar gene organisation to α_{2u}-globulin, its rat equivalent (Clark et al.,

Figure 6.4. Comparison of gene structures. Exons are shown as open boxes, the coding regions are shaded. The phase of each intron is indicated: 0, splicing between two codons; I, splicing between 1/3 (5') and 2/3 (3') of a codon; and II, splicing between 2/3 (5') and 1/3 (3') of a codon. BLG, as before; MUP, murine major urinary protein (Clark et al., 1984); α1-AGP, human α1-acid glycoprotein (Dente et al., 1987); RBP, as before; apo-D, human apolipoprotein D (Drayna et al., 1987). The size and position of the apolipoprotein D gene exon I have not been determined (*).


BLG
O
II
I
I
I
136
140
74
111
105
42
180
MUP
O
II
I
I
I
161
134
74
111
102
46
240
α1-AGP
O
II
I
I
O
154
143
71
108
104
194
RBP
O
II
I
I
60
129
137
107
4500 bp
213
288
apo-D
O
II
I
*
157
122
89
?
5600bp
2500bp
414
1kb

1984).

The BLG and MUP genes encode 162 amino-acid polypeptides whereas α1-acid glycoprotein (AGP), RBP and apolipoprotein-D (apo-D) genes encode 187, 183 and 167 amino-acid polypeptides, respectively. The BLG, MUP and AGP (Dente et al., 1985) genes contain six protein-coding exons whereas RBP has five and apo-D (Drayna et al., 1987) has four exons. In the RBP and apo-D genes transcription starts in a short non-coding exon absent from the BLG, MUP and AGP genes. In contrast, BLG and MUP have a seventh exon which is entirely non-coding.

Analogous exons are often similar in size. The first protein-coding exon ranges in size from 129 bp in the RBP gene to 161 bp in the MUP gene and encodes 32 (BLG/MUP) to 41 (apo-D) amino-acids. Exon III of apo-D is shortest at 122 bp in length, the analogous exon of AGP being 143 bp in length. Exon III is 74 bp in the BLG and MUP genes, 3 bp larger than exon III in AGP. The analogous exon in the apo-D gene is 89 bp in length and 107 bp in the RBP gene. Exon IV of the BLG and MUP genes are also identical in size and 3 bp longer than exon IV in the AGP gene. The BLG gene exon V is 3 bp larger than that of the MUP gene and 1 bp larger than that of the AGP gene. Exon V of the RBP gene is 213 bp in length, similar to the combined size of exons IV and V of the BLG and AGP genes and is identical to the combined size of exons IV and V of the MUP gene. Exons VI of the BLG, MUP, AGP and RBP genes encode the final short stretch of amino-acids. In the apo-D gene exon V contains the sequences encoded by BLG, MUP and AGP genes' exons IV, V and VI and by the RBP gene exons V and VI.

Analogous introns vary considerably in size, by as much as 78 bp to 5.6 kb (RBP/apo-D intron II). They are, with one exception, in phase with respect to the reading frame. Thus, the intron I splice junction of the BLG gene falls between two codons. This is conserved in the other four genes. The single exception is intron V of AGP. Here the splice occurs between two codons whereas in the other genes it occurs

between 1/3 (5') and 2/3(3') of a codon.

The similarity between these gene structures is good evidence for evolutionary relatedness. Many families of homologous proteins have now been described which share similar gene structures. All the kinds of differences seen in this gene family between members, have been described in other gene families. The presence in the RBP gene of a large exon of 213 bp which encodes sequences present in exons IV and V of the BLG, MUP and AGP genes suggests that there may have been an intron deletion event in the RBP gene. Similarly, the absence of yet another intron from the apo-D gene is probably due to the deletion of that intron from the apo-D gene. For example, an instance of intron loss has been described for the insulin genes. Two insulin genes are found in the rat. The rat insulin I gene contains one intron. The rat insulin II gene, in common with human, dog, guinea pig, chicken and hogfish insulin genes, contains two introns, clearly indicating intron loss from the rat insulin I gene (see Steiner et al., 1985). Although intron insertion events have been described they appear to be much rarer than intron deletions (Rogers, 1985).

The BLG and MUP genes contain a 3' non-coding exon which is absent from the other three genes. Because of little apparent sequence homology, it is difficult to assess whether the presence of intron VI is due to an intron insertion/deletion event. It is interesting to note that transcription of MUP genes gives either the seven exon form of mRNA or a second form which arises by read-through of exon VI and transcription termination in intron VI, 5' of exon VII (Clark et al., 1984). This may represent an intermediate in the evolution from a six exon gene to a seven exon gene (or vice versa), whilst not losing the old form (as suggested by Gilbert, 1978). There does not appear to be any homology between the 5' end of intron 6 of MUP and BLG genes and exon VI of the RBP gene.

As described above, introns are in phase with respect to the reading frames for the five genes. Any mutations which alter this would destroy the reading frame,

thereby preventing downstream sequences from being correctly translated. This would be deleterious and so it is perhaps not surprising that the reading frame is highly conserved. Indeed analogous exons in the five genes differ in size by multiples of three (whole codon differences). The single exception to this is the final intron of the AGP gene. However this change would not have affected many amino-acids and those present in the final exon do not form part of β-strands or α-helix (by comparison with BLG/RBP). Nevertheless, the sequences in this exon do encode a cysteine which participates in the formation of a disulphide bridge. This disulphide bridge is clearly important as it is highly conserved in many members of the superfamily (see below). Rogers (1986) reviewed similar "frameshifts" in a *Drosophila* non-fibrillar collagen gene, the TFIIIA gene and the rat fibronectin gene. The mechanisms by which this kind of "frameshift" occurs are not clear, since all possible mechanisms (so far described) would require either more than one mutational event to be applicable in these cases (thereby suggesting a stage during which the gene was not "active") or the insertion of an intron preferentially into a particular region. With respect to the former possibility, both man and mouse have two functional genes encoding AGP (Dente et al., 1987; Cooper and Papaconstantinou, 1986). Nevertheless, both copies of the gene have the same exon/intron arrangement.

6.5 A DIVERSE FAMILY OF PROTEINS

A number of other proteins are now known to be members of this family, making up the present fifteen described members. Pervaiz and Brew (1985) noted an amino-acid homology between human HC (HCHU), BLG and RBP, whilst Lee et al. (1987) noted homology between these and a protein present in the cells of the frog

Figure 6.5. DOTPLOT analysis of secretory proteins. The dotplot shows a comparison of the different proteins known to share amino-acid sequence homology with each other (also see figure 6.6). Each amino-acid sequence was enlarged by the addition of "." to the C-terminal ends, to increase all sequences up to a uniform size of 200 amino-acids for convenience. The sequences were then joined to each other so that a single comparison could be performed. The comparison (window = 20, stringency = 10) shows that each protein exhibits some sequence similarity with most other proteins in the family. A central region of similarity is seen. Some proteins exhibit greater similarity with some proteins, than with others. BLG, MUP, AGP, RBP and apo-D as before; HC, human protein HC (Kaumeyer et al., 1986; Traboni and Cortese, 1986); BG, frog Bowman's gland protein (Lee et al., 1987); ESP, rat epididymal secretory protein (Brooks et al., 1986); and PUR (Berman et al., 1987) amino-acid sequences were derived from cDNA sequences. INCYN, tobacco hornworm insecticyanin (Riley et al., 1984); C8G, human complement C8γ polypeptide (Hunt et al., 1987; Haefliger et al., 1987) and PEG, human α2-pregnancy endometrium protein (PP14) (Bell and Smith, 1988) are amino-acid sequences. The PEG amino-acid sequence is incomplete. See section 6.5 for details.

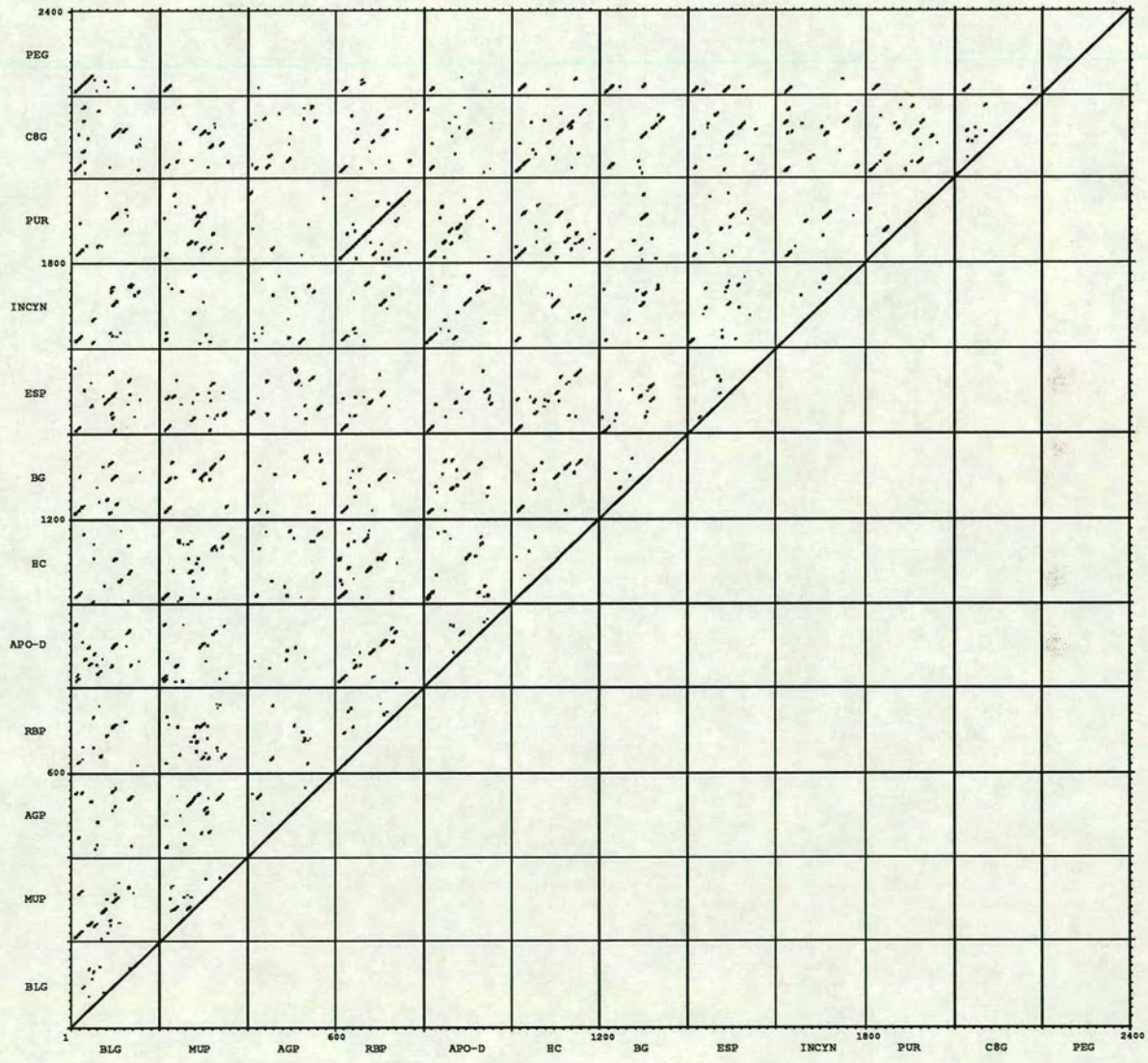

2400
PEG
C8G
PUR
1800
INCYN
ESP
BG
1200
HC
APO-D
RBP
600
AGP
MUP
BLG
1
BLG
MUP
AGP
600
RBP
APO-D
HC
1200
BG
ESP
INCYN
1800
PUR
C8G
PEG
2400

Figure 6.6. DOTPLOT analysis of secretory proteins. The dotplot shows a comparison of the different proteins known to share amino-acid sequence homology with OBP, CRN-A and CRN-C. The analysis was identical to that performed for figure 6.5. These sequences only became available after the first comparison had been done. This dotplot therefore, extends the comparison done for figure 6.5. OBP, rat odorant binding protein (Pevsner et al., 1988b); CRN-A and CRN-C, crustacyanins -A and -C (J. B. C. Findlay, personal communication).

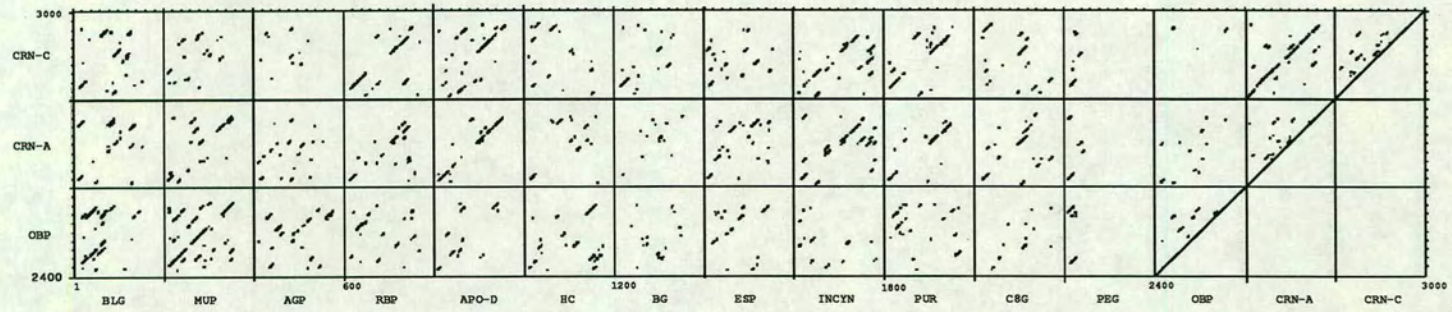

3000
CRN-C
CRN-A
OBP
2400
1
600
1200
1800
2400
3000
BLG
MUP
AGP
RBP
APO-D
HC
BG
ESP
INCYN
PUR
C8G
PEG
OBP
CRN-A
CRN-C

Figure 6.7. Alignment of amino-acid sequences. Sequence alignments were made using the UWGCG program GAP (Devereux et al., 1984) on a VAX computer. The program uses the algorithm of Needleman and Wunsch (1970). A comparison table devised by Staden (1982) based on the relatedness odds matrix MDM78 of Dayhoff (1969) was used. The table allows conservative matches to be identified, as well as perfect matches. Firstly the amino-acid sequences of the five proteins whose gene structures are known were aligned exon by exon. The gapped alignments of BLG, MUP, AGP, RBP and apo-D generated in this manner were used to align the remaining ten amino-acid sequences using GAPOUT files. Positions of exon/intron junctions (where known) are shown as straight lines where splicing occurs between two codons and as a box around an amino-acid, if splicing occurs within a codon. Numbered horizontal bars indicate particularly well conserved amino-acids. The sequences are described in the legend to figures 6.5 and 6.6. Standard one letter codes are used for the amino-acids. The "consensus" sequence shows majority matches. If two or more amino-acids are equally abundant at any position, none is shown. Imperfectly conserved matches are in lower case, universal presence is shown in upper case.

Amino-acid Sequence Alignment

olfactory organs, in the Bowman's gland (BG). Drayna et al. (1986) described a relationship between apo-D, tobacco hornworm insecticyanin (INCYN) (Riley et al., 1984), BLG, RBP, HCHU and α_{2U}-globulin (α_{2U}-G). Berman et al. (1987) cloned a chicken mRNA encoding purpurin (PUR), which is highly homologous to RBP. A complement polypeptide C8 Gamma (C8G) has homology with HCHU (Hunt et al., 1987; Haefliger et al., 1987). Pevsner et al. (1988b) have recently described a bovine olfactory protein (OBP) (which may be analogous to the frog BG ?), whilst two polypeptides which associate to form lobster retinoid-binding protein, crustacyanin-A and -C (CRN-A and -C) have been purified and sequenced (J. B. C. Findlay, personal communication). Partial amino-acid sequencing of a human placental protein, pregnancy associated endometrial α_2-globulin (PEG) has shown that it is highly homologous to BLG (Huhtala et al., 1987; Bell et al., 1987; Bell and Smith; 1988). Finally, a protein synthesised by the rat epididymis has been described (Brooks et al., 1986; Brooks, 1987). The structures of genes encoding these proteins have not been published.

The UWGCG program DOTPLOT (Devereux et al., 1984) was used to compare all amino-acid sequences to each other (suggested by A. Coulson, Univ. of Edinburgh) (figures 6.5 and 6.6). Each sequence was made up to 200 amino-acids by the addition of a tail of "." so that a uniform length is obtained. The comparison shows that there is distinct homology extending across the length of the proteins, with few large gaps. RBP and PUR are particularly closely related (as possibly are BLG and PEG). CRN-A and CRN-C are also closely related. AGP appears to share least homology with the other proteins and may be the most distantly related member of the family (see section 6.7).

Figure 6.7 shows an alignment of the amino-acid sequences of these proteins. Firstly the amino-acid sequences whose gene structures are known were

aligned exon by exon. The amino-acid sequences of the other proteins were then aligned to these. It is evident that exons encode similar motifs in the BLG and RBP genes and the conservation of the five gene structures indicates that analogous exons encode "homologous" structural units. Differences in analogous exons can be taken as evidence of a mutational event in that exon, thereby identifying the region in which a gap must be placed. It thus seems reasonable to align amino-acid sequences exon by exon, where possible.

All alignments were done using the UWGCG program GAP (Devereux et al., 1984) on a VAX operating system. It agrees well with comparisons reported by the other authors. Furthermore, the alignment obtained by comparison of the distances in space between BLG and RBP amino-acids (Lindsay Sawyer, personal communication) is in good agreement with the above alignment.

Looking at any pair of sequences in figure 6.7 shows that low amino-acid homology is spread across the entire sequence (with the exception of RBP versus PUR, which shows quite a high homology (50%)). CRN-A and -C are 59% homologous. Five regions are particularly well conserved within the family. These have been numbered *1* through *5,* in the figure. *1* is a twenty amino-acid region centred around a sequence G-X-W (glycine-X-tryptophan) (where X is any amino-acid), which is maintained perfectly in all fifteen proteins. This suggests that G-X-W plays a major role in the function of these proteins. Indeed, it has been shown that the tryptophan (W) residue is involved in binding retinol by BLG (Fugate and Song, 1980; Papiz et al., 1986). Region *3* is a second relatively large, highly conserved run of amino-acids. In particular, the run T-D-Y (threonine-aspartic acid-tyrosine) is present in eleven out of fourteen sequences (this region of PEG has not yet been sequenced). A smaller region at *4* contains the highly conserved R-X-P (arginine-X-proline).

The amino-acids at *2* and *5* are cysteine residues which are highly

conserved. The cysteine at *2* is present in ten out of fourteen proteins (the presence or absence of a cysteine in PEG at position *2* or *5* is not yet known). A cysteine is not present in apo-D, INCYN, CRN-A and CRN-C. The cysteine at *5* is conserved in all fourteen sequences (a cysteine is present at position 194 in the insecticyanin, four amino-acids away from *5*). The cysteines at *2* and *5* form a disulphide bridge in BLG (Swaisgood, 1982), RBP (Laurent et al., 1985), AGP (Schmid et al., 1974), HCHU (Lopez et al., 1981) and C8G (Haefliger et al., 1987). This suggests that the other proteins which also have these two cysteines also form this disulphide bridge between the two. Apo-D, INCYN, CRN-A and -C have a cysteine at position 53 in place of one at *2*. This cysteine participates in a disulphide bridge with the cysteine near *5* in the case of INCYN (Riley et al., 1984). This bridge is probably also present in apo-D, CRN-A and CRN-C.

The exon/intron junctions of the five genes described in section 6.4 are shown on the alignment. A vertical line indicates splicing between codons, a boxed amino-acid indicates splicing within the codon encoding that amino-acid. It can be seen that exon/intron junctions are well conserved within the alignment and thus predictions about the other gene structures, can be made. Indeed a similar alignment to that shown in figure 6.7 was obtained before the apo-D gene had been described. One exon/intron boundary proved to be absolutely correct (that between apo-D gene exons II and III). Two other junctions were present in approximately the positions that would be predicted. Similarly the HCHU gene exon I also ends as is predicted (C. Traboni, personal communication). It is, however, necessary to take care when making such predictions since it has already been shown that these genes have differences. Indeed, it could not be predicted that the apo-D gene would have only four protein-coding exons. The gene structure of the HCHU gene is also likely to differ from the rest of the family since the mRNA for HCHU encodes not only HC but another polypeptide HI-30, which is cleaved proteolytically from HCHU (Kaumeyer et al.,

1986; C. Traboni, B. Akerstrom, personal communications). However 5' exons may be expected to be similar to those for the five genes described. Structural similarities between some members of the family but not others enables possible variant gene structures to be predicted (see section 6.7).

Craik et al. (1982; 1983) proposed that exon/intron junctions map to the protein surface and that 'intron sliding" may occur, causing changes in polypeptide sizes. On examination of the exon/intron junctions in figure 6.7, it can be seen that gaps often occur around splice sites. In particular, there is a great deal of variation at the third splice (at about position 110) where β-strand F of RBP starts, unlike BLG. However, no gaps have been acquired at regions *1* and *5*.

The low amino-acid homology between family members appears not to be of great importance, since despite low homologies BLG and RBP have very similar 3D structures, the main differences being at external loops. The 3D structures of cabbage white butterfly and tobacco hornworm insecticyanins have recently been published (Huber et al. ,1987; Holden et al., 1987). INCYN forms a β-barrel made up of eight β-strands. Much of the RBP and INCYN 3D structures can be superimposed with deviations at external loops. Hence, this family of proteins probably all have similar 3D structures and similar gene structures. The conservation of 3D structures without high amino-acid homologies suggests that this β-barrel structure does not require many, specific amino-acids and/or the low homologies are due to differences in function.

Table 6.1: Properties and functions.

The table summarises some properties of the related proteins. The first column lists the abbreviated names of the proteins. In brackets are other names by which the proteins are also known. The second column shows the species in which the proteins have been characterised. In the third and fourth column are the molecular weights (in daltons) and the number of amino-acids making up the proteins, respectively. In a number of cases the actual molecular weights are given in brackets, with the molecular weights expected from amino-acid sequence also presented. In each case difference between actual and expected molecular weights is due to glycosylation. The fifth column shows whether the protein forms multimers, or not. Also listed are the tissue(s) in which the gene is expressed and the fluid(s) into which the protein product is secreted (asterisks mark presumed tissue or fluid). The eighth column lists any ligands which have been found to bind to the protein. In the case of apo-D binding to cholesterol has been postulated but not shown. Similarly, HCHU is associated with a yellow-brown ligand which has been postulated to be a retinoid. The next column shows proteins which interact with the proteins (where known). The penultimate column describes the function of each protein. See text for more information. NK = Not Known. For CRN-A and -C, little information is yet available.

The final column notes the references from which the table was compiled. These are listed here. (1) Papiz et al., 1986; (2) Futterman and Heller, 1972; (3) Fugate and Song, 1980; (4) Shaw et al., 1983; (5) Clark et al., 1984; (6) Shahan et al., 1987; (7) Vandenbergh et al., 1975; (8) Dente et al., 1985; (9) Ganguly et al., 1967; (10) Kerkay and Westphal, 1968; (11) Ricca et al., 1981; (12) Friedman, 1983; (13) Laurent et al., 1985; (14) Weech et al., 1986; (15) Drayna et al., 1986; (16) Tejler and Grubb, 1976; (17) Lopez et al., 1981; (18) Kastern et al., 1986; (19) Mendez et al., 1986; (20) Lee et al., 1987; (21) Brooks et al., 1986; (22) Brooks, 1987; (23) Riley et al., 1984; (24) Holden et al., 1987; (25) Huber et al., 1987; (26) Huhtala et al., 1987; (27) Bell et al., 1987; (28) Bell and Smith, 1988; (29) Berman et al., 1987; (30) Rao et al., 1987; (31) Ng et al., 1987; (32) Haefliger et al., 1987; (33) Pevsner et al., 1986; (34) Pevsner et al., 1988a; (35) Pevsner et al., 1988b.

Table 6.1: Properties and functions

Protein	Species	Size MW.	AA	Subunits	Tissue(s)	Fluid(s)	Ligand	Protein(s)	Function(s)	References
BLG	sheep	18,300	162	dimeric	mammary gland	milk	retinol	intestinal receptor	possible vitamin A transport to young	1, 2, 3,
MUP (α_{2u}-G)	rodents	18,700	162	monomeric	liver, salivary, lachrymal and preputial glands,	serum,urine, saliva*, tears*, seminal fluid*	NK	NK	possible binding of pheromones	4, 5, 6, 7,
AGP	man	18,900 (40,000)	187	monomeric	liver	serum	progesterone and other steroids	NK	mediates inflammatory response	8, 9, 10, 11, 12
RBP	man	22,900	183	monomeric	liver	serum	retinol	transthyretin	binds and transports retinol in the serum	13,
apo-D	man	19,300 (33,000)	169	NK	adrenal,kidney, pancreas, liver, intestine	serum, gut secretions*	cholesterol ?	lecithin:cholesterol acyltransferase	possibly involved in cholesterol transport	14, 15,
HCHU (α_1-MG)	man	20,600 (31,000)	183	some dimers	liver	serum,urine cerebrospinal fluid	yellow-brown ligand -a retinoid (?)	Immunoglobulin A	involved in mediating neutrophil chemotaxis	16, 17, 18, 19
BG	frog	20,300	160	monomeric	olfactory	nasal mucus* epithelium	odorants	neuronal receptors	probably involved in presentation of odorants to receptors on neurons	20,
ESP	mouse	18,500	165	monomeric	epididymis luminal fluid*	epididymal	NK	sperm mebrane protein(s)	binds sperm membrane	21, 22
INCYN	moth	21,400	189	tetrameric	larval fat body?	hemolymph	biliverdins;	NK	involved in colouration, photoprotection, oxygen radical quenching	23, 24, 25
PEG (PP14)	man	? (28,000)	?	dimeric	placenta, secretory endometrium	amniotic fluid	NK	NK	synthesised during mid- to late luteal and early pregnancy stages. Function ?	26, 27, 28,
PUR	chicken	21,900	175	monomeric	photoreceptor cells of retina	interphotoreceptor cell matrix	retinol	NK	present in retina; promotes embryonic neural adhesion and aids survival	29,
C8G	man	20,300 (22,000)	182	monomeric	liver	serum	NK	C8α, C8β (and other complement proteins?)	part of complement C8	30, 31, 32,
OBP	rat	18,100	172	monomeric	nasal epithelium	mucus	odorants	neuronal receptors	presentation of odorants to neuronal receptors	33,34, 35,
CRN-A) > CRN-C)	lobster	18,900 20,100	171 178	tetrameric	NK	NK	a retinoid	NK	NK	J. B. C. Findlay, personal communication.

6.6 PROPERTIES AND FUNCTIONS

The above analysis shows that the fifteen proteins probably arose from the same ancestral gene present before the divergence of arthropod and chordate lines. They appear to share similar gene and protein structures. Therefore, at least some of the properties and functions of these proteins should be conserved. A number of proteins in the family, including BLG, do not have defined functions. In this section I discuss some characteristics of the various proteins and suggest that potential roles for BLG can be speculated.

In table 6.1 a number of characteristics and functions of family members have been tabulated. Family members are present in arthropods (INCYN and CRN-A and -C), amphibians (BG), birds (PUR) and mammals. Some of the proteins characterised in mammals are involved in carrying out defined important functions and their presence may be expected in other classes or phyla. Certainly, Berman et al. (1987) have cloned a chicken RBP cDNA so a serum retinol transporting protein is presumably widespread in chordates. Other proteins appear to be more specific. For example, BLG is a milk protein and is therefore mammal-specific. Even within mammals it is absent from man and rodents.

The proteins have similar-sized, small subunits of 160-189 amino-acids, with molecular weights of 18-23 kdal. AGP, apo-D, HCHU and PEG are highly glycosylated; their actual molecular weights are 40,000, 30,000 31,000 and 28,000 (Wagh et al., 1969; Weech et al., 1986; Tejler and Grubb, 1976; Bell et al. 1987) respectively. A number of the proteins are multimeric; BLG is dimeric in milk, PEG is dimeric and INCYN forms a tetramer. Crustacyanin is a tetramer of two CRN-A and two CRN-C subunits. The other proteins seem to be monomeric.

All the proteins are synthesised in one or more secretory tissues and in each

case are secreted into a body fluid. MUP, AGP, RBP, apo-D, HCHU and C8G are synthesised in the liver and secreted into blood serum; BLG is synthesised in the mammary gland, BG and OBP are synthesised in the olfactory epithelium.

Many are known to bind small, lipophilic molecules. A large range of such molecules are bound. AGP binds steroids, RBP binds retinol, apo-D probably binds cholesterol and INCYN binds a biliverdin pigment molecule. The probable similarity of 3D structures and the absolute conservation of amino-acids shown to be important for ligand binding (the region containing G-X-W; the tryptophan interacts with retinol bound by BLG (Fugate and Song, 1980; Papiz et al., 1986) strongly suggest that all fifteen members bind such small molecules.

Another conserved feature of these proteins appears to be their interaction with other proteins. Some form multimers (BLG, INCYN and crustacyanin). Some associate with other serum proteins. RBP associates with the serum protein transthyretin. apo-D associates with lecithin:cholesterol acyltransferase whilst HC interacts with immunoglobulin A. C8G is non-covalently associated with the complement C8β polypeptide and is covalently disulphide-bonded to C8α.

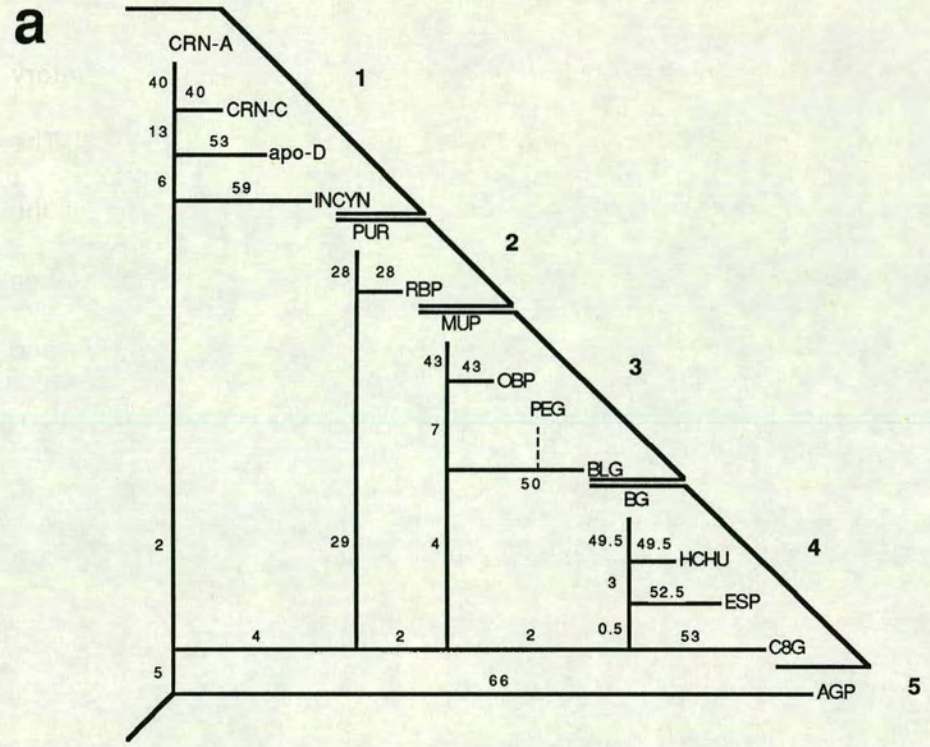
Members of the family which transport lipophilic molecules often associate with cell surface receptors. RBP gives up its retinol to a receptor in the retina. BG and OBP presumably interact with neuronal receptors. ESP binds to a sperm membrane protein and BLG may interact with a receptor present in the microvilli of calf intestine.

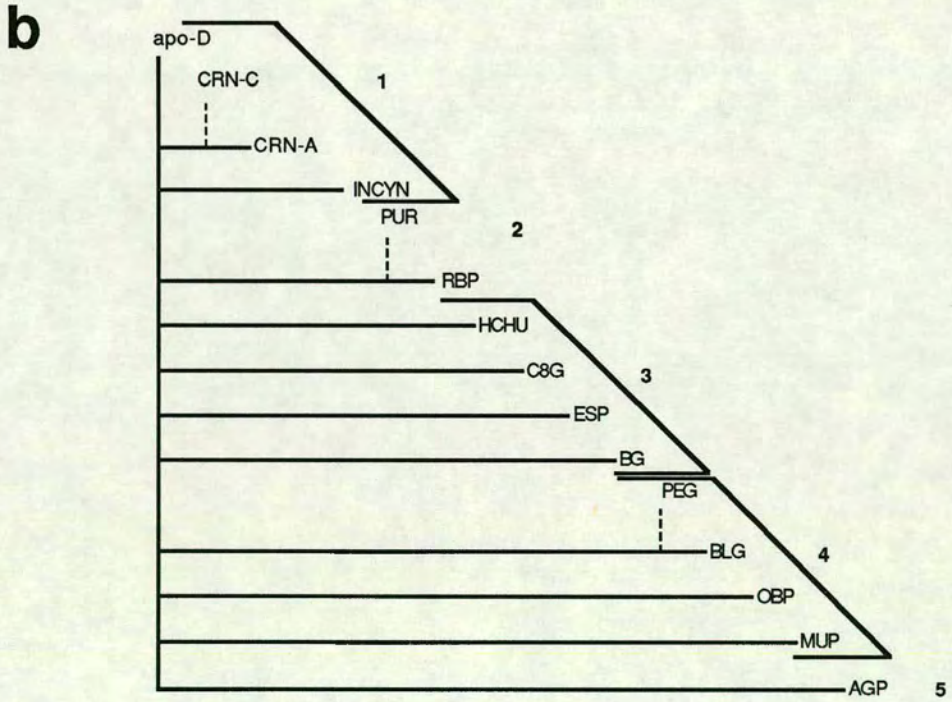
Table 6.2

	A	C	OBP	BLG	MUP	AGP	RBP	apo	HC	BG	ESP	INC	PUR	C8G	PEG
CRN-A	000	080	125	121	112	145	120	103	120	125	124	121	111	129	
CRN-C		000	126	122	123	155	129	109	135	119	122	131	122	133	
OBP			000	108	087	121	120	127	110	115	117	121	115	117	
BLG				000	095	125	108	126	115	108	110	116	108	106	
MUP					000	114	111	122	098	113	104	118	110	107	
AGP						000	153	137	144	129	131	156	149	138	
RBP							000	106	128	108	110	125	055	114	
apo-D								000	115	121	116	108	107	132	
HCHU									000	099	107	133	117	107	
BG										000	104	119	116	105	
ESP											000	123	116	107	
INCYN												000	125	129	
PUR													000	117	
C8G														000	
(PEG	038	042	044	022	038	045	036	037	036	036	040	039	035	037	000)

The table shows the distance matrix data obtained for pair-wise comparisons of the fifteen proteins. It lists the number of differences between pairs of proteins. The UWGCG program DISTANCE was used to work out the number of matches between pairs of sequences. These were substracted from the size of the smaller of the two proteins compared in each case. The distance data obtained for PEG is shown for comparison but was not used for generating KITSCH trees. Fifty six amino-acids of PEG sequence are known (see figure 6.7). Some of the protein names have been further abbreviated (A=CRN-A, C=CRN-C, apo=apo-D, HC=HCHU and INC=INCYN).

Figure 6.8. Phylogenetic trees from amino-acid sequences of the secretory proteins. (a) shows a rooted distance tree based on the data in table 6.2. The lengths of the internodes and branches are shown. The program assumes that all the proteins are contemporaneous, having arisen from one ancestor. Thus, the distance from the tree root to any tip is the same. Distances are indicated by numbers and add up to the same total from the root to any tip. (b) shows an unrooted parsimony tree based on amino-acid differences (figure 6.7). The internodes and branch lengths have no significance. For both analyses 50 replicates were done on 5 arrangements of the input sequence information (a total of 250 replicates). In each case the same tree was generated for (a). For (b) 96% of the trees were identical to the one shown. The rest gave trees with branching, indicating a more closer relationship than is implied by the tree in (b). In these cases the relationships were more similar to those in (a). The broken lines indicate proteins which were not used in the tree building analysis but whose amino-acid sequence similarity clearly indicates relationship to another protein (for example RBP and PUR). From the distance tree in (a), five groups are evident. These are also shown in (b).


a
CRN-A
40
40
CRN-C
13
53
apo-D
6
59
INCYN
PUR
1
2
28
28
RBP
MUP
43
43
OBP
3
PEG
7
BLG
50
BG
2
29
4
49.5
49.5
HCHU
4
3
52.5
ESP
0.5
53
C8G
4
2
2
5
66
AGP
5


b
apo-D
CRN-C
1
CRN-A
INCYN
PUR
2
RBP
HCHU
C8G
3
ESP
BG
PEG
BLG
4
OBP
MUP
AGP
5

6.7 ORIGIN AND DIVERGENCE OF THE FAMILY

A number of methods have been developed for comparing related DNA and protein sequences for degree of evolutionary relatedness. The various algorithms described are generally "parsimony" or "distance" methods for determining phylogenies. Distance methods use differences between sequences, derived by pairwise comparisons, to obtain a tree(s) and assess it for goodness of fit by statistical means. Parsimony methods find tree(s) requiring the fewest nucleotide changes to explain evolution of the observed sequences. Parsimony methods use the entire sequences for deriving relationships whereas distance methods only use the overall difference between sequences (Felsenstein, 1981; 1985; 1988).

Distance matrix data and parsimony programs for inferring phylogenies, distributed by J. Felsenstein (PHYLIP 3.0, copyright University of Washington) were used to compare the fifteen proteins. Program KITSCH analyses distance matrix data by the method of Fitch and Margoliash (1967). An assumption made is that the data-set represents contemporaneous species which have diverged equally from the root of the tree. Thus the total length from the root of the tree to any species will be the same. An evolutionary clock is used, thereby giving rooted trees. The distance matrix data were derived using the UWGCG program DISTANCE (see table 6.2).

Program PROTPARS was used for parsimonious analysis of the sequences. PROTPARS gives an unrooted tree, making no assumptions about evolutionary rates and using no out-groups. PROTPARS determines the least number of changes required to get from sequence A to B using the method of Eck and Dayhoff (1966) and Fitch (1971), which counts the number of nucleotide changes needed for the protein sequence changes. Changes need to be consistent with the genetic code so that some changes are allowed only via an intermediate change (two changes) whereas others

require only one change. The trees generated by these methods are shown in figure 6.8. Both programs generate trees which may be dependent on the order in which the sequences are entered. Thus replicates, in which the order is changed, need to be carried out in order to confirm that the best tree is being generated (see legend to figure 6.8).

All replicates of KITSCH yielded the same tree (figure 6.8a). The tree shows that the proteins in this family fall into five groups. One group consists of the proteins CRN-A, CRN-C, apo-D and INCYN (group *1*). A second group contains RBP and PUR. The third group comprises MUP, OBP and BLG whilst the fourth group contains BG, HCHU, ESP and C8G. AGP appears to be only distantly related to the family. The PROTPARS generated tree (figure 6.8b) shows many of the same features of the other tree. It also differs in many respects from the KITSCH-generated tree. The proteins shown to be most closely related by KITSCH are seen to be most similar. Thus proteins in groups *1* to *4* of the first tree cluster together in the PROTPARS-generated trees. However, they do not form clear groupings (with the possible exception of group *1*). The reason for this may be the different kinds of analyses carried out by the two programs. Whereas KITSCH looks only at the the number of overall differences in pair-wise comparisons, PROTPARS compares the entire sequences. Nevertheless, the trees generated by PROTPARS were consistent with those generated by KITSCH. It may, therefore, be possible to draw some valid conclusions from the trees.

The possibility that the group *1* proteins may indeed be more closely related to each other than to the other proteins is also suggested by examination of a conserved structural feature. The cysteine at *5* (in figure 6.7) is probably involved in all group *1* proteins in the formation of a disulphide bridge different from that found in the other proteins (see section 6.5). Similarly HCHU, C8G, BG and ESP form group *4*. Comparison of C8G amino-acid sequence with the other proteins gives the

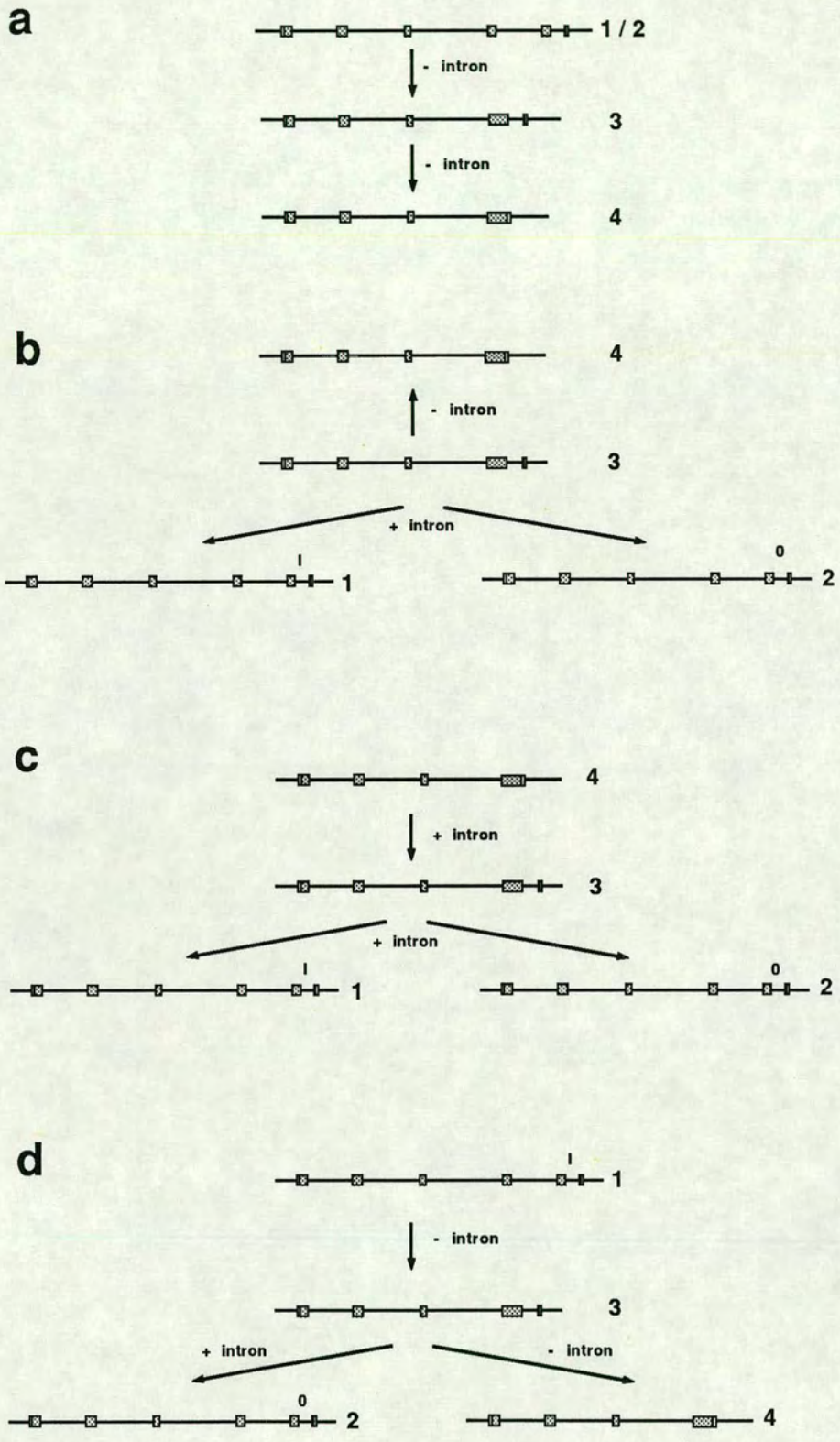
best score against HCHU (Hunt et al., 1987; Haefliger et al., 1987). Comparison of the cysteines present shows that HCHU, C8G and BG contain three cysteines, two of which are involved in formation of the described disulphide bridge. The third cysteine of HCHU and C8G (at position 47 in the alignment) forms a disulphide bridge with Immunoglobulin A and C8α, respectively. There is no evidence for inter-molecular disulphide bonds between any of the other proteins. BG could be involved in a similar disulphide bond formation. However, this cysteine is absent from ESP.

The group *1* proteins appear to be related to the other family members through RBP and PUR. BG, HCHU, ESP and C8G may form another group of proteins which are more similar to each other than to the other proteins. A fourth possible group contains OBP, MUP and BLG (and probably PEG). It is interesting to note that the two olfactory proteins, BG and OBP belong to different groups, in these analyses, possibly having different origins within the family. AGP appears to be highly diverged from all of the other proteins. The parsimony trees generated suggest that it may be linked through MUP. In contrast a comparison of AGP, α_{2u}-globulin, BLG, RBP and HCHU, using the program RELATE (Dayhoff, 1978; Dayhoff et al., 1983) (at 250 PAMs) suggested that AGP may be more closely related to HCHU than to MUP (J. O. Bishop, personal communication).

As was noted above, consideration of the findings from the trees with information about protein structures suggests that CRN-A, CRN-C, INCYN and apo-D are indeed more closely related to each other than to other family members. However, few other structural features can be directly compared. Gene structure differences and similarities should reflect these putative relationships. Gene structures of one or more proteins in group *1*, *2* and *3* have been determined but those of HCHU, BG, ESP or C8G genes have not been described.

The gene structures have been compared in section 6.4. It is clear from the

Figure 6.9. Possible routes for evolution of the genes. Fewest events leading to the evolution of the present genes are considered and four possible routes are shown. It has been assumed that any event is rare and the same event will not occur more than once. Only protein-coding exons have been considered. Open boxes represent exons, shading indicates protein coding regions. BLG/MUP-like (1), AGP-like (2), RBP-like (3) and apo-D-like (4) genes are labelled.


a
1 / 2
- intron
3
- intron
4
b
4
- intron
3
+ intron
I
1
0
2
c
4
+ intron
3
+ intron
I
1
0
2
d
I
1
- intron
3
+ intron
- intron
0
2
4

analysis that the BLG and MUP gene structures are very similar. Both genes contain six protein-coding exons and a seventh non-coding exon. The AGP gene structure resembles that of the BLG/MUP genes but an intron "frameshift" is found. The RBP and apo-D genes contain a small exon encoding 5' untranslated sequences and both genes have one less intron. In addition the apo-D gene has one intron less than the RBP gene. If it is assumed that the same intron insertion/deletion events would be unlikely to occur more than once, it is possible to speculate a number of possible routes that may have led to the present day genes; requiring the fewest number of events (as these events are likely to be rare) (figure 6.9). Only protein-coding exons are shown in the figure since these exons must be greatly restricted in the kinds of changes that can occur without affecting proteins sequence (as discussed above). In figure 6.9a the ancestral gene is assumed to be BLG/MUP/AGP-like (*1/2*), containing six protein-coding exons. Deletion of an intron would generate an RBP-like gene (*3*) which could generate an apo-D-like gene (*4*) by the loss of a further intron. It is also possible that a six-exon gene may have undergone loss of two introns to directly generate an apo-D-like gene. A separate intron deletion would have given an RBP-like gene. The RBP and apo-D genes contain a short first exon encoding 5' untranslated sequences, which is absent from the BLG/MUP, AGP and HCHU (C. Traboni, personal communication) genes (groups *2*, *3*, *4* and *5*). The fact that these four genes contain no 5' untranslated exon suggests that the ancestral gene giving rise to the RBP-like gene (3) may not have had such an exon. This would suggest that the RBP-like gene generated as shown in figure 6.9a may have acquired a short 5' exon before further duplication occurred; rather than this event occurring twice independently. One of these genes then lost an intron to yield the apo-D ancestral gene (*4*).

Three other possible evolutionary routes are shown in figure 6.9b-d. All these would require both intron deletion and intron insertion events to have taken place. As mentioned above, intron insertion have been observed much more rarely

than intron deletions in gene families. Nevertheless, the three routes present possible ways in which the apparent intron "frameshift" seen in the AGP gene relative to the BLG/MUP and RBP genes, may have arisen.

Whatever the paths of gene evolution followed, the differences in gene structures correlate well with the groupings generated by tree building. Thus BLG and MUP, in group *3*, have very similar gene structures. The BLG/MUP genes show some differences to the RBP (group *2*), apo-D (group *1*) and AGP genes. The apo-D gene appears to be more similar to the RBP gene than to the BLG/MUP/AGP genes, as discussed above.

Unfortunately, little information is available regarding the genes in group *4*. It is nevertheless clear from preliminary evidence that the HCHU gene is different from the other groups' gene structures. Like the other genes it contains an identically placed intron (see above) and amino-acid sequence comparisons suggest that it may also share other exon/intron similarities (see figure 6.7). Its cDNA sequence shows, however, that the gene encodes another polypeptide, named HI-30. HI-30 is a 42 kdalton polypeptide which forms part of a 180 kdal protein, inter-α-trypsin-inhibitor (ITI). ITI is a serine protease inhibitor found in human serum. HI-30 is proteolytically cleaved by trypsin *in vitro* and contains all of the anti-proteolytic activity of the 180 kdal form (Kaumeyer et al., 1986; and references therein). ITI has been described as being a single polypeptide. Cloning of a human liver cDNA which encodes only the 40 kdal polypeptide suggests that this may not be so. Preliminary experiments by C. Traboni suggest that the HCHU gene locus also codes for the large subunit of ITI as a distinct mature mRNA. This may correlate with the findings of Bourguignon et al. (1983) who reported multiple mRNAs in baboon liver code for polypeptides related to ITI. Thus the HCHU gene will apparently contain some structural differences from all the genes so far described.

Data available on the gene encoding C8G suggests that it may be similar to

that encoding HCHU, at least in its complexity. C8G is disulphide bonded to C8α in the serum. C8α-C8G is non-covalently associated with C8β. This complex interacts with the larger complement complex containing C5β-C9. Genetic data suggests that C8α-C8G and C8β are encoded by two separate genes (Alper et al., 1983), that is C8α-C8G is the product of a single gene. Distinct mRNAs encoding C8α, C8β and C8G have recently been cloned by the same lab (Rao et al., 1987; Howard et al., 1987; Ng et al., 1987, respectively). In these papers the authors present Northern data for RNA from HepG2, baboon liver and rat liver, which show that each cDNA probe hybridises to an RNA species of the expected size. On the evidence of this data they conclude that C8α and C8G are encoded by two separate genes. They do show, however, that the C8α cDNA hybridises to two RNA species (2.5 kb and 1.5 kb in length - the expected size is 2.5 kb). HepG2 RNA was not probed with the C8G cDNA. Baboon liver RNA probed with the C8G cDNA gives a band of about 1.0 kb. There is some hybridisation to an RNA of about 2.5 kb (the size obtained for the C8α mRNA)and some larger RNA species. These may be pre-mRNAs. It is possible, however, that the C8G gene acts analogously to the HCHU gene. Indeed there appears to be some amino-acid sequence similarity between about twenty amino-acids surrounding the site of proteolytic cleavage of HCHU-HI-30 and the C-terminal amino-acid sequences of C8G and the N-terminal amino-acid sequences of C8α (the HCHU sequences are present N-terminal of HI-30 sequences).

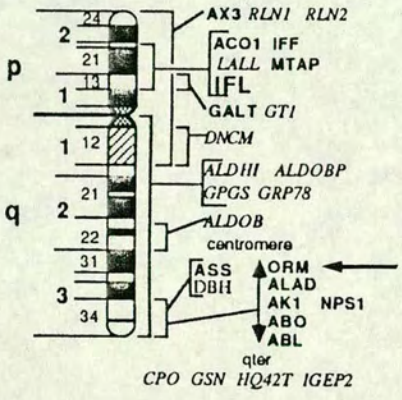
Although it is difficult to ascertain whether evolutionary trees generated for a set of proteins as highly diverged as these are real, it seems that the relationships seen correlate well with differences in protein and gene structures between family members. Cysteine residues are clearly important for the structures of these proteins as seen by their high conservation. They are characteristically conserved in the five groups and the presence of different cysteine and disulphide bridges in the groups correlates very well with the tree data. Inspection of gene structures, like

protein structures, shows that all members have similar conserved features and some differences. Members of the five groups differ in some aspect from genes of other groups. More gene structures need to be determined before the evolutionary pathways can be more clearly elucidated. Nevertheless BLG and MUP genes share almost identical gene structures and apparently belong in the same group. Thus, the trees generated by distance matrix and parsimony analyses may reflect the routes by which these proteins diverged from each other.

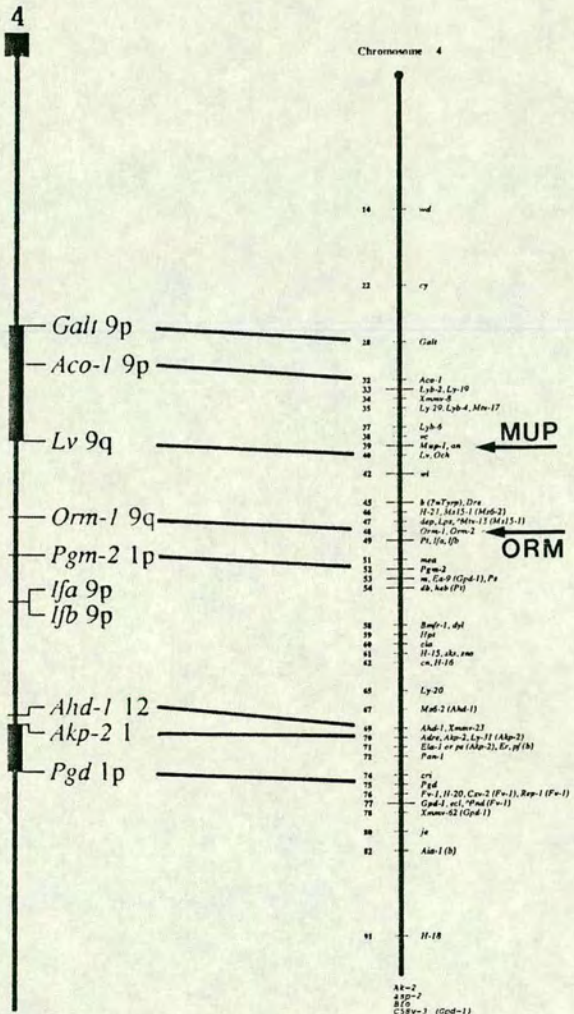
In a number of gene families, some or all members show linkage, for example, the MHC class I and II genes (see Klein, 1986), all six γ-crystallin and some β-crystallin genes (see Wistow and Piatigorsky, 1988) and all casein genes (Gupta et al., 1982; Gaye et al., 1986). The chromosomal location of the human AGP, RBP, apo-D, HC and C8G genes and the mouse AGP and MUP genes, have been determined. AGP and MUP genes have been mapped to mouse chromosome 4 (Bennett et al., 1982; Bishop et al., 1982; Krauter et al., 1982; Searle et al., 1987). The two human AGP genes have been mapped by *in situ* hybridisation, to chromosome 9q31-q34.1 (Webb et al., 1988). Traboni et al. (1987) placed the human HC gene on chromosome 9. They have found that the HC gene maps to 9q31-q33 (C. Traboni, personal communication). Thus the AGP, HC (α1-microglobulin) and MUP genes are probably linked on mouse chromosome 4, since the human AGP and HC genes are linked in man. On the other hand, the human apo-D gene maps to p14.2➡qter of chromosome 3 (Drayna et al., 1987). Using human-hamster cell hybrids the RBP gene has been assigned to human chromosome 10 (Rocchi et al., 1987). In addition, the human C8α–γ locus has been placed on chromosome 1p36.2-p22.1 (Rogde et al., 1987; see Morton and Bruns, 1987).

It is interesting that the AGP and HC genes are linked and the C8G gene, which is much more closely related to the HC gene than the AGP gene (Hunt et al.,

Figure 6.10. Chromosomal location. Human chromosome 9 and mouse chromosome 4 are shown. The middle map shows the chromosomal positions of mapped human loci which map to mouse chromosome 4. Standard abbreviations for loci are used (see Human Gene Mapping 9 (1987): Ninth International Workshop on human gene mapping. Cytogenet. Cell. Genet. vol. 46). The human chromosome 9 map is reproduced from McKusick (1986). The two mouse chromosome 4 maps were provided by Dr. J. T. Eppig (The Jackson Laboratory). ORM is AGP.



9



Ak-2 1p
Eno-1 1p
Fuca 1p
Lck 1p
Pnd 1p
Xpa 1q
Tsha 6q
Mos 8q
Ggtb-1 9p

1987; Haefliger et al., 1987; this chapter), is present on a different chromosome, in man. As figure 6.10 shows, however, many mouse chromosome 4 genes map to human chromosome 1, whilst many others including the AGP genes, map to human chromosome 9. It is possible that a translocation event has separated the C8G and HC genes in man, and it is also possible that the two genes are linked in mouse. Nevertheless, some of the proteins in three groups (*3*, *4* and *5*) show linkage. This may indicate that these three groups are more closely related to each other than to group *1* and *2* proteins. However, the presence of the C8G gene on human chromosome 1 and not chromosome 9 indicates that strong conclusions can not be made from linkage data. Nevertheless, it may also be possible to investigate orthology. For example, BLG, MUP and PEG may be orthologous, since two, or all three proteins have not been found in any one species. BLG is almost certainly absent from rodent and human milks. Chromosomal localisation may allow their relationship to be further clarified.

6.8 MEMBERS OF A LARGER GROUP OF PROTEINS?

Another large group of proteins some of which transport small lipophilic molecules intracellularly, have been described. These proteins are about 134 amino-acids in length and have molecular weights of about 14 kdal. Approximately eight different proteins have been isolated and include retinol-binding proteins, retinoic-acid-binding proteins, fatty-acid-binding proteins, as well as others whose function is not clear, such as peripheral myelin P2 protein (Takahashi et al., 1982; Sacchetini et al., 1986; Demmer et al., 1987; Jones et al., 1988; and references therein). The proteins contain a G-X-W (glycine-X-tryptophan) motif at the

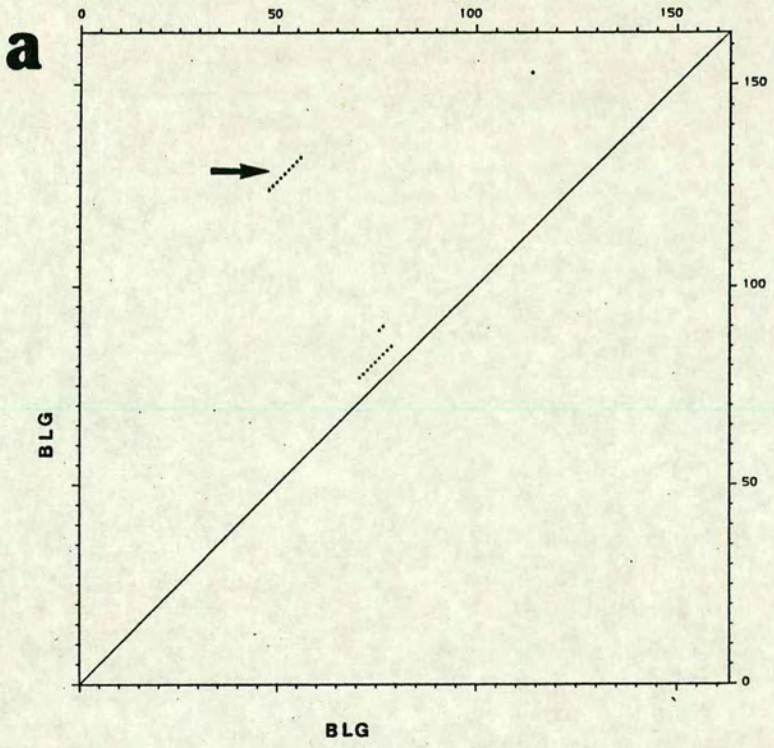
N-terminal end. It seems possible that the larger secretory proteins are related to these smaller, intracellular proteins and suggests a possible "relay" system of related proteins, for the transport of hydrophobic (lipophilic) molecules through aqueous body and cell fluids.

Amino-acid sequence comparisons between RBP and cellular retinol-binding protein II (cRBPII) shows a region of possible homology stretching over about 30 amino-acids (figure 6.11b). No real sequence similarity is seen outside this region (except G-X-W at the N-terminal region).

Examination of the amino-acid sequence of RBP has led to speculation about a possible internal duplication (Rask et al., 1981). This duplication presumably must have occurred before the divergence leading to these fifteen proteins since they are all similar in size. Also, Kolde and Braunitzer (1983b) showed internal sequence similarity in the sheep BLG amino-acid sequence. DOTPLOT comparison of BLG shows such a region of internal similarity (figure 6.11a). No similar region was clear in MUP, AGP or apo-D proteins (not shown). Figure 6.12 shows part of the alignment shown in figure 6.7. Exon/intron junction positions are arrowed. Regions of possible internal sequence similarity are underlined for RBP and BLG (from Rask et al., 1981; Kolde and Braunitzer, 1983b; figure 6.11a). Figure 6.12b shows an alignment of these regions. Also shown is an alignment of cRBPII with RBP over the region obtained in the DOTPLOT in figure 6.11b (figure 6.12c).

The region of RBP which appears to have been duplicated is about 35 amino-acids in length. This is also the approximate difference in size between RBP and cRBPII if the first protein-coding exon of RBP is not counted (this exon contains the signal peptide and the most N-terminal sequences - a signal peptide is not required in intracellular proteins). It is therefore, possible that the larger proteins arose by an internal duplication within the smaller proteins and the acquisition of sequences which form the signal peptide. In the case of BLG a smaller 20 amino-acid

Figure 6.11. DOTPLOT analysis of BLG and RBP. (a) shows the dotplot comparison of BLG amino-acid sequence versus itself. This shows that at least one region of internal homology (arrowed) may exist (amino-acids 45-65 show some sequence similarity to amino-acids 120-140). Window = 30, stringency = 6.0. (b) shows a dotplot comparison of RBP versus cellular RBPII (cRBPII - Demmer et al., 1987). Amino-acids 1-18 of RBP are signal peptide sequences. This shows a region of sequence similarity over about 25 amino-acids. A second, shorter region of similarity is found. This may correspond to part of the region of internal duplication reported by Rask et al., 1981. Window = 20, stringency = 6.0.


a
0
50
100
150
150
100
50
0
BLG
BLG

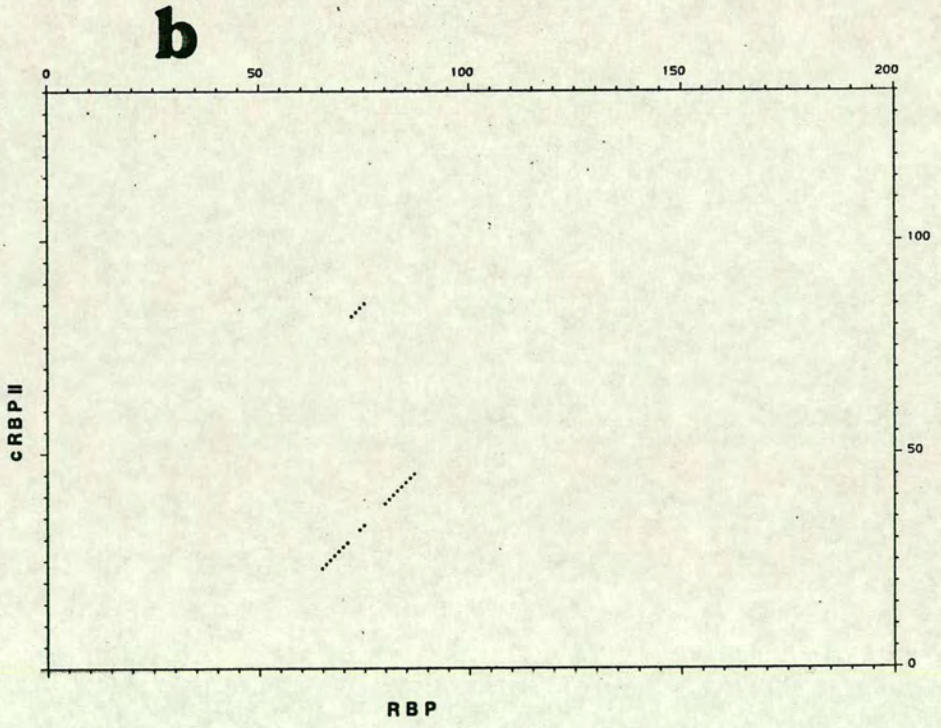

b
0
50
100
150
200
100
50
0
cRBPII
RBP

Figure 6.12. Internal homology in the protein sequences. (a) shows the amino-acid sequence alignment of BLG, MUP, AGP, RBP and apo-D shown in figure 6.7. Exon/intron junctions are indicated by arrows. The partially boxed sequences indicate regions which show internal sequence similarity. These internal homologies were suggested by Rask et al. (1981), for RBP and by Kolde and Braunitzer (1983b), for BLG. The BLG sequences underlined show sequence similarity in the dotplot analysis in figure 6.10a. (b) shows the regions of internal homology described by Rask et al. (1981) and Kolde and Braunitzer (1983b). In (c) sequence similarity between RBP and cRBPII is shown (see figure 6.10b), as determined by Demmer et al. (1987).

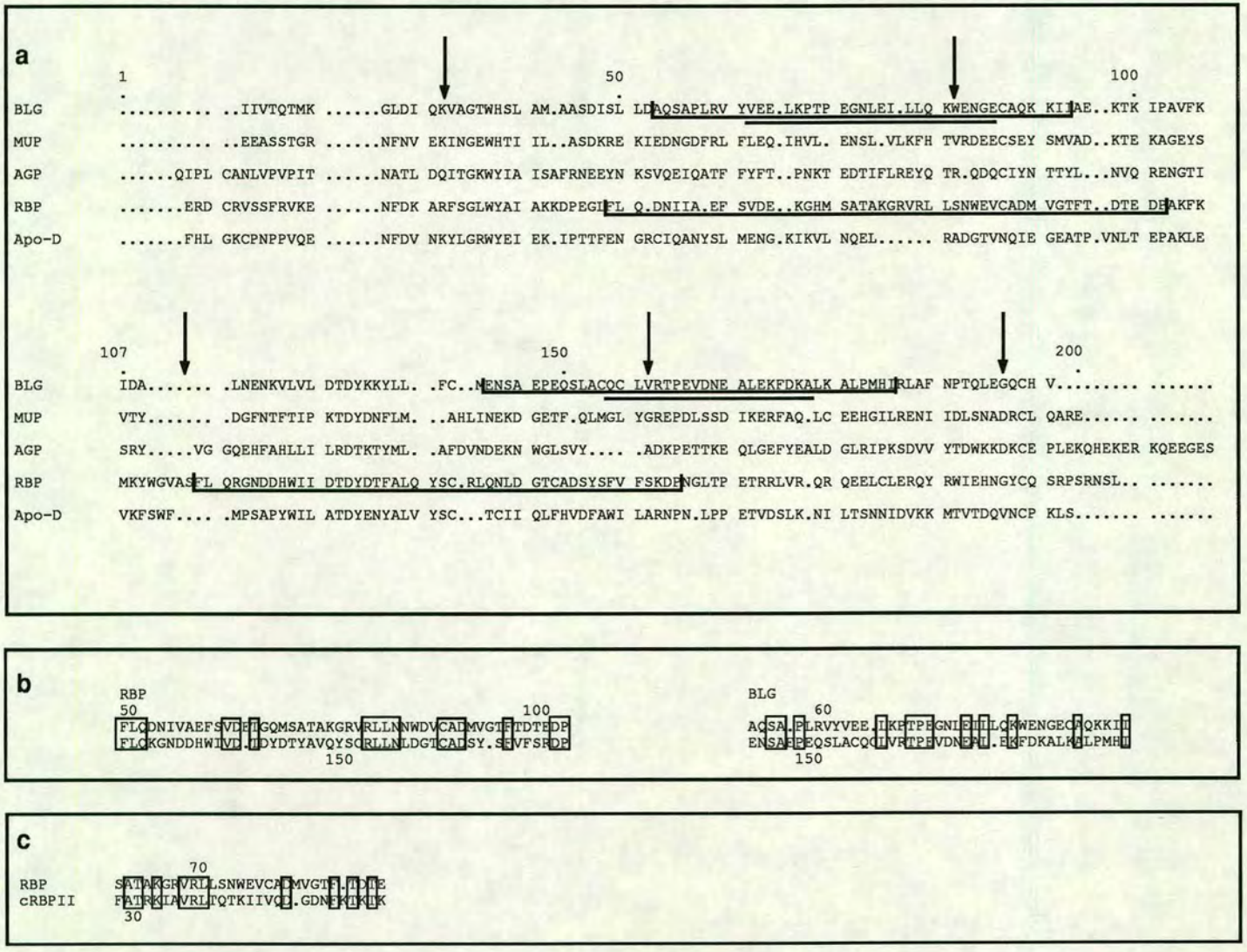
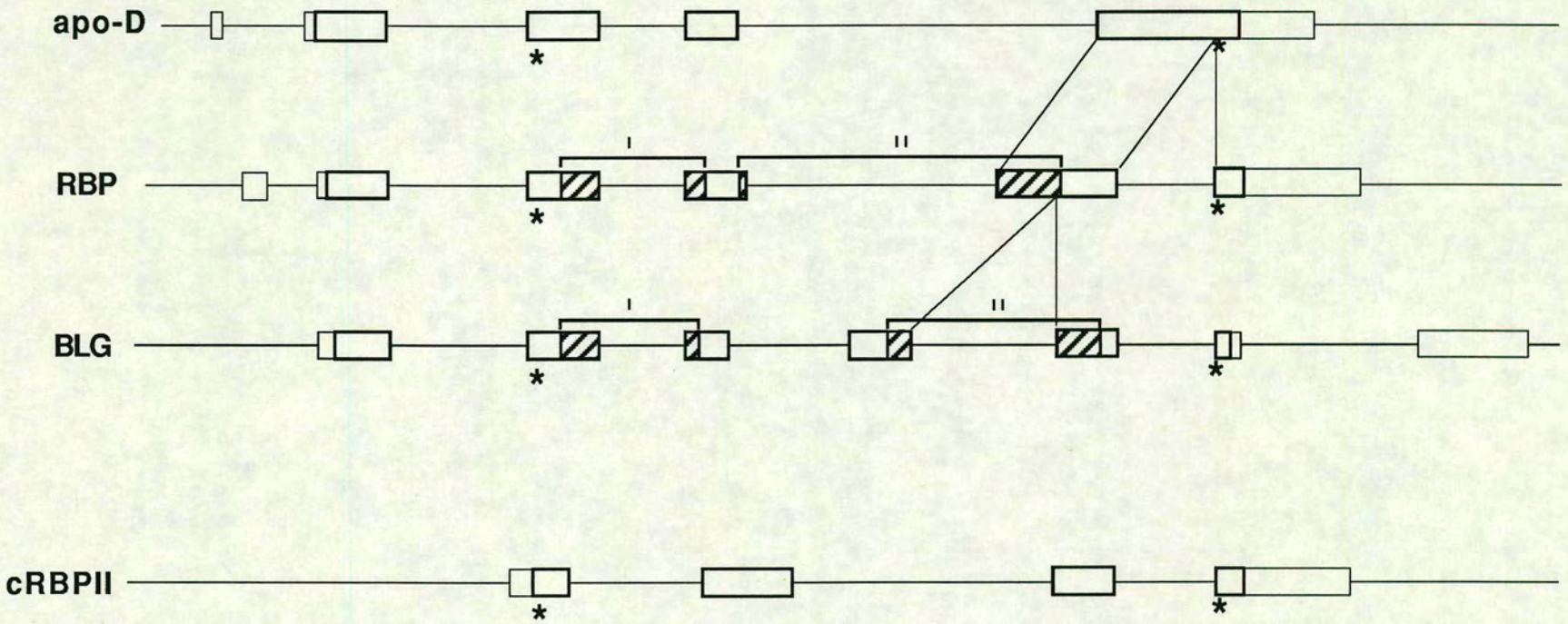

a
1
50
100
BLG ....IIVTQTMK ......GLDI QKVAGTWHSL AM.AASDISL LDAQSAPLRV YVEE.LKPTP EGNLEI.LLQ KWENGECAQK KIIAE..KTK IPAVFK
MUP ....EEASSTGR ......NFNV EKINGEWHTI IL..ASDKRE KIEDNGDFRL FLEQ.IHVL. ENSL.VLKFH TVRDEECSEY SMVAD..KTE KAGEYS
AGP ......QIPL CANLVPVPIT ......NATL DQITGKWYIA ISAFRNEEYN KSVQEIQATF FYFT..PNKT EDTIFLREYQ TR.QDQCIYN TTYL...NVQ RENGTI
RBP ....ERD CRVSSFRVKE ......NFDK ARFSGLWYAI AKKDPEGLFL Q.DNIIA.EF SVDE..KGHM SATAKGRVRL LSNWEVCADM VGTFT..DTE DPAKFK
Apo-D ....FHL GKCPNPPVQE ......NFDV NKYLGRWYEI EK.IPTTFEN GRCIQANYSL MENG.KIKVL NQEL..... RADGTVNQIE GEATP.VNLT EPAKLE
107
150
200
BLG IDA..... .LNENKVLVL DTDYKKYLL. .FC..MENSA EPEQSLACQC LVRTPEVDNE ALEKFDKALK ALPMHIRLAF NPTQLEGQCH V..... .....
MUP VTY..... .DGFNTFTIP KTDYDNFLM. ..AHLINEKD GETF.QLMGL YGREPDLSSD IKERFAQ.LC EEHGILRENI IDLSNADRCL QARE..... .....
AGP SRY.....VG GQEHFAHLLI LRDTKTYML. .AFDVNDEKN WGLSVY.... .ADKPETTKE QLGEFYEALD GLRIPKSDVV YTDWKKDKCE PLEKQHEKER KQEEGES
RBP MKYWGVASFL QRGNDDHWII DTDYDTFALQ YSC.RLQNLD GTCADSYSFV FSKDENGLTP ETRRLVR.QR QEELCLERQY RWIEHNGYCQ SRPSRNSL.. .....
Apo-D VKFSWF.... .MPSAPYWIL ATDYENYALV YSC...TCII QLFHVDFAWI LARNPN.LPP ETVDSLK.NI LTSNNIDVKK MTVTDQVNCP KLS..... .....
b
RBP
50
100
FLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDP
FLCKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSY.SFVFSRDP
150
BLG
60
AQSAPLRVYVEELKPTPEGNLEILLQKWENGECAQKKII
ENSAEPEQSLACQCLVRTPEVDNEAL.EKFDKALKALPMHI
150
c
70
RBP SATAKGRVRLLSNWEVCADMVGTF.TDTE
cRBPII FATRKIAVRLLTQTKIIVQL.GDNEKTKIK
30

Figure 6.13. Comparison of gene structures. The figure shows the gene structures of BLG (six protein-coding exons as in the MUP and AGP genes), RBP and apo-D. The regions of possible internal homologies are also shown (hatched). Protein-coding regions are shown as heavy boxes;

noncoding, exonic sequences as faint boxes. The gene structure of cRBPII (Demmer et al., 1987) is shown below. The asterisk under the second protein-coding exons of apo-D, RBP and BLG denotes the conserved G-X-W. The asterisk under the final protein coding exon is one of the conserved cysteines. These two sequences are present in the first and last protein-coding exons of the cRBPII gene.


apo-D
RBP
BLG
cRBPII
I
II
I
II

region with possible sequence similarity was found (Kolde and Braunitzer, 1983b).

The schematised gene structures of BLG, RBP and apo-D (typifying the different gene structures in figure 6.4) are shown in figure 6.13, together with the cRBPII gene structure (the gene structures of rat cRBPII, rat liver fatty-acid-binding protein and mouse aP2 have been described and are very similar (Demmer et al., 1987)). Also shown are the regions of internal homology in RBP.

Close to the start of the second protein-coding exons of BLG, RBP and apo-D genes and close to the translation initiation codon of cRBPII is the motif G-X-W, present in all members of the secretory protein family and the intracellular protein family (with the exception of rat liver fatty-acid-binding protein which contains G-X-Y (Y=tyrosine)). A cysteine is present in the final exon of the cRBPII gene in a similar position to the one in BLG and RBP genes although it is probably not involved in disulphide bond formation (Jones et al., 1988). As stated above, the putative regions of duplication in RBP and BLG genes would have increased the size of the smaller, intracellular proteins to that of the secretory proteins (not including the first protein-coding exon). Region I spans parts of RBP and BLG second and third protein-coding exons. The part of RBP and BLG second protein-coding exon not included in region I is about the same size as the cRBPII exon I protein-coding part. It is possible that duplication of region II could have led to the present RBP and BLG second protein-coding exons.

It is thus possible that the first and last exons of the cRBPII gene are related to the second and last protein-coding exons of the BLG and RBP genes. If this were the case then possible routes of evolution could be envisaged. Certainly it would indicate that an apo-D-like gene would probably not have been the ancestral gene for the secretory protein family (figure 6.9, mechanism C), since an intron is present in an analogous position to the one in the BLG and RBP genes.

It is not clear that the two gene families are related, or whether an internal

duplication occurred at any time in the evolution of the secretory proteins. Indeed, although region I (figure 6.12a and 6.13) of BLG and RBP is similarly placed, region II is clearly different and is encoded by different exons. Nevertheless, the presence of G-X-W and a cysteine at similar positions in the mature proteins of the two families appears to indicate a possible relationship between the two. The recently published three-dimensional structures of rat intestinal fatty-acid binding protein and bovine P2 myelin protein show that both proteins form a β-barrel composed of 10 anti-parallel β-strands with a non-polar pocket as in RBP, BLG and INCYN. Some superimposition of the myelin P2 three-dimensional structure, with that of RBP can be done but there are many major differences between the two structures. The similarities may however, be indicative of evolutionary relatedness; although convergent evolution due to similarity in function is also possible. The independent evolution of G-X-W in a similar position in both families is, however, statistically unlikely since other proteins which bind similar ligands do not contain this motif. Whether the two families are related or not the importance of the amino-acids G-X-W cannot be ignored.

6.9 SUMMARY

6.9.1 THE LIPOSECS

A large family of proteins has been described. Some of these are known to share protein and gene structure similarities. Despite low amino-acid homology regions of high conservation are present. This suggests that all family members probably have similar β-barrel protein structures. Protein parsimony and distance matrix analyses indicate that the proteins fall into five groups. This is partially

confirmed by the apparent similarities and differences in gene structures. Furthermore, it is possible that these proteins are related to a large family of intracellular proteins many of which carry out functions similar to the larger, secretory proteins. Indeed, it is possible to imagine how duplication of intracellular carriers of hydrophobic molecules such as retinol, was followed by the acquisition of sequences which could act as signal peptide sequences allowing secretion of the proteins into fluids.

Many of the proteins are known to bind small lipophilic molecules and conservation of sequences known to be involved in ligand-binding is strongly indicative of conserved function. Thus all members probably bind such molecules (some members of each tree-generated group are known to bind lipophilic molecules). Whether their sole function is to act as carriers of these molecules is as yet unclear. Certainly three of the proteins, AGP, HCHU and C8G, are involved in immune responses and whether their role is to deliver a particular ligand(s) to the site of immune action has yet to be established. It was noted above that the odorant-binding proteins BG and OBP appear not to be directly related. This may imply convergent evolution of proteins in different groups within the family, or that these proteins have diverged greatly and have given rise to the other proteins in their groups.

Doolittle (1981) noted that "gene duplication often begets more gene duplication" and this appears to be clear in the family described here where presumably one ancestral gene has given rise to a large, diverse family of proteins. Furthermore, a number of the proteins are themselves part of multigene families. There are at least two AGP genes in man and rodents, possibly more than one BLG gene in sheep and there are about 35 MUP (and α_{2u}-globulin) genes. The MUP genes are interesting particularly because a unit of two genes (a "pseudogene" and a functional

gene) has been duplicated a number of times (Ghazal et al., 1985).

Pervaiz and Brew (1987) suggested that the family should be called "lipocalins" because they bind lipophilic molecules within a flower-like "calyx". This name can also be applied to the intracellular proteins so I suggest the name "liposecs" to describe the lipophile-binding secretory proteins. Lipocalins would then be an appropriate name for both families of proteins.

6.9.2 Implications for the function of BLG

As discussed in this chapter, it now seems highly likely that BLG binds and transports a lipophilic molecule from the mother to the young. It is unlikely to have lost such a function during evolution since all the short amino-acid segments conserved in the other proteins are present, none having been lost. Furthermore, BLG has been shown to bind retinol, although retinol transport is probably not its real function. The presence of a gut receptor in cows also suggests that it still functions in transporting a lipophilic molecule. However, it may do so whilst associated with immunoglobulins or caseins (with which it has been linked), since many of the other proteins also interact with various proteins, including immunoglobulins. In light of the information presented here I do not believe that BLG is involved in phosphate metabolism, as has been suggested (Thompson and Farrell, 1974).

It also appears that the human endometrial protein PEG, and BLG are more closely related to each other than to the other proteins. It is possible that the two proteins may be involved in transporting the same ligand during pregnancy PEG) and lactation (BLG). It has also been suggested that PEG may have replaced BLG (or vice versa) in some species. It has been noted (L. Sawyer, personal communication) that

BLG is absent from the milks of mammals with haemo-chorial (e.g. man) or haemo-endothelial (e.g. rodents) placentas. These placental types allow close contact between fetal and maternal blood supplies unlike the more "primitive" placental types which have more extensive barriers between maternal and fetal blood (e.g. see van Tienhoven, 1968). Some mammals, having one of the three placental types, named epithelio-chorial, syndesmo-chorial and endothelio-chorial, have been found to produce BLG in their milks (e.g. pigs, dogs and sheep, respectively). BLG has also been found in marsupial (Eastern Grey Kangaroo - Godovac-Zimmerman, 1987) and possibly in monotreme (L. Sawyer, personal communication) milks, in both of which there is no placental development. Thus absence of BLG may be correlated with placental type. Sawyer suggests that the mammals with placental types where the maternal and fetal bloods come into closest contact may not need BLG to provide a ligand during lactation, but may be able to use endometrially secreted PEG to carry out the same function earlier in the life of the young. BLG would not be required in these mammals and this may explain its absence from these mammals. It remains to be seen whether any other mammals (i.e. rodents) have a PEG gene.

ADDENDUM

The sequence of a cDNA for the human placental protein PEG (PP14) has recently been published by Julkunen et al. (1988). The partial amino-acid sequence has already been published and is included above. The deduced amino-acid sequence of PP14 consists of a putative signal peptide of 18 amino-acids and a 162 amino-acid mature polypeptide. The mature polypeptide has a predicted molecular weight of 18,787 daltons. The actual protein is glycosylated with an actual molecular weight of about 28,000 daltons (Bell et al., 1987). Figure A1a shows a GAP comparison (Devereux et al., 1984) of the amino-acid sequences of ovine BLG and PP14. Both proteins contain 162 amino-acids. The alignment shows that the two sequences have 42.6% sequence identity, no gaps being introduced. The four cysteines in PP14 correspond to the four cysteines which in bovine BLG form intramolecular disulphide bridges. Figure A1b shows a comparison of BLG amino-acid sequences from a number of mammals, including the Eastern Grey Kangaroo (see Godovac-Zimmermann, 1988 for references). The comparison shows that the divergence between BLG from the various species and PP14 is not much greater than that between BLGs from different species. This indicates that PP14 is probably orthologous to BLG. Its divergence from other BLGs appears to reflect the divergence of man from these mammals. This suggests that PP14 may bind the same lipophilic molecule(s) that BLG does. Both proteins could be involved in the transport of this molecule to the young at different stages of its development (PP14 in amniotic fluid, BLG in milk) (see section 6.9.2).

Southern blot analysis of human genomic DNA shows that the PP14 gene is greater than 20 kb in size, or more likely from the gene structure evidence presented in chapter 6, there are multiple PP14-related gene sequences in the human genome. Julkunen et al. (1988) found no hybridisation to MCF-7 cell RNA (a human breast

cancer cell line, which expresses some milk protein genes). This may mean that PP14, or related genes, are not expressed in the mammary gland.

Figure A1. Comparison of BLG and PP14. (a) GAP (Devereux et al., 1984) alignment was performed on the ovine BLG (top) and PP14 (lower) (Julkunen et al., 1988) amino-acid sequences. 42.6% amino-acid sequence identity was found, no gaps being introduced. The standard single letter codes are used to denote the amino-acids. The cysteines in BLG involved in intramolecular disulphide bridge formation are asterisked. (b) shows pairwise comparisons of BLG amino-acid sequences from eight species (taken from Godovac-Zimmermann, 1988). Also shown is a pairwise comparison between PP14 and the BLGs. The numbers in bold lettering across the diagonal from top left to bottom right show the number of amino-acids which make up the mature proteins. Above this diagonal are the number of amino-acid substitutions in each pairwise alignment. The numbers below the diagonal show percentage of amino-acid substitutions.

a

Ovine BLG v PP14

```
  1 IIVTQTMKGLDIQKVAGTWHSLAMAASDISLLDAQSAPLRVYVEELKPTPEGNLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYK 100
      ||   |   | ||||| |||   |||     |||||    | |||| |||| |  |||  |  ||   |||  |  ||       |    |||||
  1 MDIPQTKQDLELPKLAGTWHSMAMATNNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWENNSCVEKKVLGEKTGNPKKFKINYTVANEATLLDTDYD 100
                                                              *

                                                                                                          
101 KYLLFCMENSAEPEQSLACQCLVRTPEVDNEALEKFDKALKALPMHIRLAFNPTQLEGQCHV 162
        |  ||     | ||  || | |    | |    |  |   || |     | |  |
101 NFLFLCLQDTTTPIQSMMCQYLARVLVEDDEIMQGFIRAFRPLPRHLWYLLDLKQMEEPCRF 162
       *           *                                        *
```

b

	Bovine	Buffalo	Goat	Ovine	Horse I	Horse II	Porcine	Kangaroo	PP14
Bovine	**162**	4	8	8	68	90	62	113	94
Water buffalo	2.4	**162**	4	4	70	91	63	111	94
Goat	4.8	2.4	**162**	1	70	91	66	110	94
Ovine	4.8	2.4	0.6	**162**	70	91	66	109	93
Horse I	40.9	42.1	42.1	42.2	**162**	52	80	125	97
Horse II	54.2	54.8	54.8	54.8	30.7	**166**	101	135	96
Porcine	37.3	37.9	39.7	39.7	48.1	60.8	**158**	120	96
E. G. Kangaroo	68	66.8	66.2	65.6	75.3	78.9	72.2	**154**	120
PP14	58.0	58.0	58.0	57.4	59.9	57.8	60.8	74.1	**162**

No. of amino-acid substitutions

% of amino-acid substitutions

Chapter 7 SUMMARY

The mammary gland undergoes rapid development and differentiation during pregnancy and lactation. This development is principally regulated by oestrogen, progesterone and prolactin. Other hormones and local interactions are also important. The mammary gland's function is to provide the newborn with all their nutritional requirements. Milk contains thousands of different components, but a small number of these make up the bulk of this mammalian secretion. A number of proteins are abundant in milk and carry out specific functions, in addition to providing the substantive source of nitrogen in the diet of the young. The abundant, mammary-specific milk proteins are encoded by single copy genes and are hormonally regulated. Their expression is transcriptionally and post-transcriptionally regulated. As such these genes offer good model systems for studying temporal and tissue-specific expression. This is particularly interesting since the mammary gland is one of a few tissues which undergoes a major part of its development during adult life.

With the development of technology for making transgenic animals it has become possible to follow regulation of genes and regulatory sequences in the whole animal. Thus, tissue-specificity and temporal regulation of the milk protein genes can be analysed, and important regulatory sequences dissected. Furthermore, the absence so far of reliable mammary cell culture systems for these purposes makes the use of transgenic animals attractive. Transgenic technology can be used for directing expression of foreign genes to the mammary gland, under the control of milk protein gene promoters. Medically important proteins can be harvested from the milk (see Lathe et al., 1986; Clark et al., 1987).

In order to work on these projects the gene encoding ovine β-lactoglobulin was cloned in this lab. β-lactoglobulin is absent from mouse milk so transgenic mouse milk can easily be analysed for the presence of this protein, without problems often caused by the presence of similar endogenous genes. Work was also initiated to make transgenic sheep. Ovine milk protein gene sequences were to be used to direct human factor IX and human α_1-antitrypsin gene expression to the mammary gland of transgenic sheep.

A sheep spleen genomic DNA library was constructed and six phages were selected by hybridisation to a cDNA for ovine BLG (Gaye et al., 1986). Four of these clones were characterised by restriction mapping and hybridisation to specific 5' and 3' fragments of the cDNA (A. J. Clark's results). Initial DNA sequencing of exon I and a part of exon II by A. J. Clark, confirmed that one of the phages (SS1) contains BLG gene sequences. The entire SS1 16.2 kb genomic insert and a 10.5 kb subclone were injected into mice. Transgenic mice showed abundant levels of expression, secreting large amounts of BLG in milk, thus confirming that SS1 is a functional gene, encoding ovine BLG (Simons et al., 1987).

This thesis describes the detailed characterisation and DNA sequencing of the ovine BLG gene SS1. It was shown that the BLG gene is contained within a 4.9 kb transcription unit. The BLG gene contains seven exons, translation starts in exon I and ends in exon VI. The sequence of the entire transcription unit has been determined and 1.9 kb of 3' flanking region has also been sequenced. This shows that the gene contains sequences which agree with the consensus sequences derived from work with other genes. Thus, a potential TATA box, splice junctions and potential splice branch sites, and a polyadenylation signal, are present.

Work carried out by S. Harris, A. Archibald and others in the lab (unpublished results), has shown that 810 bp of 5' flanking sequences are sufficient

for efficient, mammary-specific expression of the ovine BLG gene in transgenic mice. The sequences from -810 to +30 were determined by S. Anderson and others (see Harris et al., 1988). In this thesis computer analysis of these 5' flanking sequences for the presence of putative transcriptional regulatory sequences, has been presented. The region from about -410 to the transcription start site contains at least four domains which may be important for transcriptional regulation of the BLG gene. No good matches to the 35 bp sequence motif described by Hall et al. (1987) were found in the appropriate region upstream of the transcription start site. Lubon and Hennighausen (1988) showed *in vitro* binding of a mammary nuclear protein to part of this sequence in the rat α-lactalbumin gene promoter. Purified HeLa cell NF1 also bound to this sequence. Good matches to potential NF1 binding sites are described in chapter 4.

In addition, ovine repeats have been mapped to SS1 and SS12. DNA sequence comparisons of one of these repeats showed that it is similar to previously described ruminant Alu-like repeats. Furthermore, analysis of the DNA sequence has shown that the ovine BLG gene is extremely G+C-rich (~60%) over the entire gene. This G+C-rich domain extends over at least 13 kb. The CpG content (and TpG + CpA) indicates that no part of this gene forms a CpG island (as defined by Bird, 1986; 1987). The significance of this unusual sequence composition is not clear. Nevertheless, such G+C-rich chromosomal domains have been described (see Bernardi et al., 1985).

The exonic structure and sequence of the BLG gene SS12 has been determined. This work shows that SS1 and SS12 encode ovine BLG gene alleles B and A, respectively. The existence of at least five haplotypes has been demonstrated.

Analysis of the methylation status of the ovine BLG gene during pregnancy and early lactation has been initiated. *HhaI* digestion of mammary and liver DNA

showed that the ovine BLG gene is undermethylated in the mammary gland, relative to liver DNA. This is in agreement with previous observations showing that genes are often undermethylated in tissues expressing a gene, relative to non-expressing tissues (see Razin and Cedar, 1984; Bird, 1986, 1987). More work is needed to describe fully the methylation status of the BLG gene.

Other work has investigated the tissue-specficity of the BLG gene and its patterns of expression in sheep during pregnancy and lactation have been analysed. BLG gene expression has been compared with histological analysis of mammary gland development and serum progesterone and prolactin levels during pregnancy and early lactation. BLG gene expression has been compared with expression of other ovine milk protein genes (α-lactalbumin and all four caseins). This analysis showed that the BLG gene is expressed at low levels in the virgin ewe, but mRNA levels remain low during the first half of pregnancy. BLG mRNA levels rose between days 90-100 of pregnancy and continued to rise into lactation. α_{s1}- and β-casein gene expression was similar to that of the BLG gene. α-lactalbumin and κ-casein mRNA levels rose from day 145 of pregnancy, whilst α_{s2}-casein expression started after parturition. This shows the the milk protein genes are differentially expressed, although some show coordinate expression. The large increase in mRNA levels of all proteins may be due to the rapid fall in progesterone levels at parturition, or because of increase in prolactin levels.

Work by S. Harris (unpublished results) has shown that in the BLG gene transgenic mice (the 16.2 kb SS1 construct) low level of BLG mRNA are present. Levels increased from day 10-12 of pregnancy and continued to rise into early lactation. The time course of expression of the BLG gene in transgenic mice was similar to that of the endogenous mouse β-casein gene. The endogenous WAP gene showed little expression until day 14-16. This suggests that the ovine BLG gene

shows similar regulation in transgenic mice, as it does in sheep. Expression of the BLG gene is coordinated with expression of the β-casein gene in both sheep and in transgenic mice. These results indicate that the BLG gene is "correctly" regulated in transgenic mice and indicate that 3.9 kb of 5' flanking, the 4.9 kb transcription unit and 7.3 kb of 3' flanking sequences are sufficient for correct tissue-specific and temporal regulation of the ovine BLG gene. This suggests that the transgenic model system will enable dissection of the sequences required for BLG gene regulation. As stated above, sequences contained within -810 to +30 are sufficient for mammary specific expression of the BLG gene in transgenic mice. Preliminary evidence suggests that transgenic mice containing only the sequences described in chapter 4 show similar patterns of expression in transgenic mice, as the transgenic mice containing the 16.2 kb construct (S. Harris, unpublished results). Further deletional analysis using transgenic mice is in progress in the lab.

The major milk proteins are not only a nitrogen source for the young. They also carry out important functions in milk. α-lactalbumin is required for lactose synthesis, the caseins keep large amounts of calcium in suspension. However, the function of β-lactoglobulin is unclear. Chapter 6 describes the detailed analysis of the relationship between BLG and many other secretory proteins present in insects, amphibians, birds and mammals, in a variety of body fluids. These proteins share similar three dimensional structures (Papiz et al., 1986; Holden et al., 1987; Huber et al., 1987). Comparison of their gene structures (where known) and the BLG gene structure showed that they are similar. Where known, the proteins bind small hydrophobic molecules. Tree building programs (using the programs devised and distributed by J. Felsenstein, Dept. of Genetics, Univ. Washington, Seattle), amino-acid sequence comparisons and gene structure comparisons showed that BLG is most closely related to rodent urinary proteins and to a human placental protein

(PEG). The detailed comparisons of these proteins presented here offers strong evidence for function being the transport of small, lipophilic molecules in body fluids.

REFERENCES

Ali, S. and Clark, A. J. (1988). Characterization of the gene encoding ovine Beta-lactoglobulin: Similarity to the genes for Retinol Binding Protein and Other Secretory Proteins. J. Mol. Biol. 199, 415-426.

Allison, L. A., Moyle, M., Shales, M. and Ingles, C. J. (1985). Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. Cell 42, 599-610.

Alper, C. A., Marcus, D., Raum, D., Peterson, B. H. and Spira, T. J. (1983). Genetic polymorphism in C8 β-chains: evidence for two unlinked genetic loci for the eighth component of human complement (C8). J. Clin. Invest. 72, 1526-1531.

Al-Shawi, R., Ghazal, P., Clark, A. J. and Bishop, J. O. (1989). Intraspecfic evolution of a gene family coding for urinary proteins. J. Mol. Evol. *In Press.*

Anderson, R. R. (1974). Endocrinological control. In, Lactation, a comprehensive treatise volume I: the mammary gland/development and maintenance. Edited by B. L. Larson and Smith, V. R. Published by Academic Press, New York and London.

Anderson, R. R. (1975). Mammary gland growth in sheep. J. Anim. Sci. 41, 118-123.

Andres, A.-C., Schonenberger, C.-A., Groner, B., Hennighausen, L., LeMeur, M. and Gerlinger, P. (1987). Ha-ras oncogene expression directed by a milk protein gene promoter: tissue specificity, hormonal regulation and tumour induction in transgenic

mice. Proc. Natl. Acad. Sci. USA 84, 1299-1303.

Angel, P., Imagawa, M., Chiu, R., Stein, B., Imbra, R. J., Rahmsdorf, H. J., Jonat, C., Herrlich, P. and Karin, M. (1987). Phorbol ester-inducible genes contain a common *cis*-element recognised by a TPA-modulated trans-acting factor. Cell 49, 729-739.

Antonarkis, S. E., Beeham, C. D., Giardina, P. J. and Kazazian, H. H. Jr. (1982). Non-random association of polymorphic restriction sites in the β-globin gene cluster. Proc. Natl. Acad. Sci. USA. 79, 137-141.

Ball, R. K., Friis, R. R., Schoenenberger, C. A., Doppler, W. and Groner, B. (1988). Prolactin regulation of β-casein gene expression and a cytosolic 120-kd protein in a cloned mouse mammary epithelial cell line. EMBO J. 7, 2089-2095.

Banerjee, M. R. (1976). Responses of mammary cells to hormones. Int. Rev. Cytology 47, 1-97.

Banerji, J. Olson, L. and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell 33, 729-740.

Barberis, A. Superti-Furga, G. and Busslinger, M. (1987). Mutually exclusive interaction of the CCAAT-binding factor and of a displacement protein with overlapping sequences of a histone gene promoter. Cell 50, 347-359.

Beato, M., Arnemann, J., Chalepakis, G., Slater, E. and Willmann, T. (1987). Gene

regulation by steroid hormones. J. Steroid Biochem. 27, 9-14.

Beckmann, J. S., Kashi, Y., Hallerman, E. M., Nave, A. and Soller, M. (1986). Restriction fragment length polymorphism among Israeli Holstein-Friesian dairy bulls. Animal Genetics 17, 25-38.

Bell, K. and McKenzie, H. A. (1964). β-lactoglobulins. Nature 204, 1275-1279.

Bell, K. and McKenzie, H. A. (1967). The whey proteins of ovine milk: β-lactoglobulins A and B. Biochim. biophys. Acta 147, 123-134.

Bell, K., McKenzie, H. A. and Shaw, D. C. (1968). Amino-acid composition and peptide maps of β-lactoglobulin variants. Biochim. Biophys. Acta 154, 284-294.

Bell, S. C., Keyte, J. W. and Waites, G. T. (1987) Pregnancy-associated endometrial α_2-globulin, the major secretory protein of the luteal phase and first trimester pregnancy endometrium, is not glycosylated prolactin but related to β-lactoglobulins. J. Clin. Endocrinol. Metabol. 65, 1067-1071.

Bell, S. C. and Smith, S. (1988). The endometrium as a paracrine organ. In, Contemporary obstetrics and gynaecology. Ed. G. V. P. Chamberlain. Butterworths Scientific Ltd., London, pp 273-298.

Bennett, K. L., Lalley, P. A., Barth, R. K. and Hastie, N. D. (1982). Mapping the structural genes coding for the major urinary proteins in the mouse: combined use of recombinant inbred strains and somatic cell hybrids. Proc. Natl. Acad. Sci. USA 79,

1220-1224.

Benoist, C. and Chambon, P. (1981). *In vivo* sequence requirements of the SV40 early promoter region. Nature 290, 304-310.

Berman, P., Gray, P., Chen, E., Keyser, K., Ehrlich, D., Karten, H., LaCorbiere, M., Esch, F. and Schubert, D. (1987) Sequence analysis, cellular localisation and expression of a neuroretina adhesion and cell survival molecule. Cell 51, 135-142.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. Science 228, 953-958.

Bienz, M. and Pelham, H. R. B. (1986). Heat shock regulatory elements function as an inducible enhancer in the *Xenopus* hsp70 gene and when linked to a heterologous promoter. Cell 45, 753-760.

Biggin, M. D., Gibson, T .J. and Hong, G. F. (1983). Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA 80, 3963-3965.

Bird, A. P. (1986). CpG islands and the function of DNA methylation. Nature 321, 209-213.

Bird, A. P. (1987). CpG islands as gene markers in the vertabrate nucleus. Trends In genetics 3, 342-347.

Birnboim, H. C. and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Research 7, 1513-1523.

Bisbee, C. A. and Rosen, J. M. (1987). DNA sequence elements regulating casein gene expression. Transcriptional Control Mechanisms, 313-323.

Bishop, J. O., Clark, A. J., Clissold, P. M., Hainey, S. and Francke, U. (1982). Two main groups of mouse major urinary protein genes, both largely located on mouse chromosome 4. EMBO J. 1, 615-620.

Bissell, M. J. (1981). The differentiated state of normal and malignant cells or how to define a 'normal' cell in culture. Int. Rev. Cytol. 70, 27-100.

Blake, C. C. F. (1978). Do genes-in-pieces imply proteins-in-pieces? Nature 273, 267.

Blake, C. C. F. (1985). Exons and the evolution of proteins. International Review of Cytology 93, 149-185.

Bolivar, F. and Backman, K. (1979). Plasmids of *Escherichia coli* as cloning vectors. Method Enzymol. 68, 245-267.

Bolander, F. F. Jr. and Topper, Y. J. (1979). Stimulation of lactose synthetase activity and casein synthesis in mouse mammary explants by estradiol. Endocrinology 106, 490-495.

Bonsing, J. and Mackinlay, A. G. (1987). Recent studies on nucleotide sequences encoding the caseins. J. Dairy Research 54, 447-461.

Bourguignon,J., Vercaigne, D., Sesboue, R., Martin, J. P. and Salier, J. P. (1983). Inter-alpha-trypsin inhibitor (ITI): two distinct mRNAs in baboon liver argue for a discrete synthesis of ITI and ITI derivatives. FEBS Lett. 162, 379-383.

Boutin, J-M., Jolicoeur, C., Okamura, H., Gagnon, J., Edery, M., Shirota, M., Banville, D., Dusanter-Fourt, I., Djiane, J. and Kelly, P. A. (1988). Cloning and expression of the rat prolactin receptor, a member of the growth hormone/prolactin receptor gene family. Cell 53, 69-77.

Breathnach, R. and Chambon, P. (1981). Organization and expression of eucaryotic split genes coding for proteins. Ann. Rev. Biochem. 50, 349-383.

Brew, K., Vanaman, T. C. and Hill, R. L. (1967). Comparison of the amino-acid sequence of bovine α-lactalbumin and hen's egg white lysozyme. J. Biol. Chem. 242, 3747-3749.

Brew, K., Castellino, F. J., Vanaman, T. C. and Hill, R. L. (1970) The complete amino-acid sequence of bovine α-lactalbumin. J. Biol. Chem. 245, 4570-4582.

Brooks, D. E., Means, A. R., Wright, E. J., Singh, S. P. and Tiver, K. K. (1986) Molecular cloning of the cDNA for two major androgen-dependent secretory proteins of 18.5 kdaltons synthesised by the rat epididymis. J. Biol. Chem. 261, 4956-4961.

Brooks, D. E. (1987) The major androgen-regulated secretory protein of the rat epididymis bears sequence homology with members of the α_{2u}-globulin superfamily. Biochem. International 14, 235-240.

Brown, P., Spooner, R. L. and Clark, A. J. (1989). Cloning and characterisation of a *BoLa* class I cDNA clone. Immunogenetics 29, 58-60.

Bucher, P. and Trifonov, E. N. (1986). Compilation and analysis od eukaryotic Pol II promoter sequences. Nucleic Acids Research 14, 10009-10026.

Burditt, L. J., Parker, D., Craig, R. K., Getova, T. and Campbell, P. N. (1981). Differential expression of α-lactalbumin and casein genes during the onset of lactation in the guinea-pig mammary gland. Biochem. J. 194, 999-1006.

Butler., J. E. (1974). In Lactation III. Eds. B. L. Larson and V. R. Smith. Publ. Academic Press, New York.

Campbell, S. M., Rosen, J. M., Hennighausen, L. G., Strech-Jurk, U. and Sippel, A. E. (1984). Comparison of the whey acidic protein genes of the rat and mouse. Nucleic Acids Research 12, 8685-8697.

Caskey, C. T. (1987). Disease diagnosis by recombinant DNA methods. Science 236, 1223-1229.

Ceriani, R. L. (1970a). Fetal mammary gland differentiation *in vitro* in response to

hormones I. Morphological findings. Developmental Biology 21, 506-529.

Ceriani, R. L. (1970b). Fetal mammary gland differentiation *in vitro* in response to hormones II. Biochemical findings. Developmental Biology 21, 530-546.

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry 18, 5294-5299.

Chodosh, L. A., Baldwin, A. S., Carthew, R. W. and Sharp, P. A. (1988). Human CCAAT-binding proteins have heterologous subunits. Cell 53, 11-24.

Church, G. M. and Gilbert, W. (1984). Genomic sequencing. Proc. Natl. Acad. Sci. USA 81, 1991-1995.

Clark, A. J. Clissold, P. M., Al Shawi, R., Beattie, P. and Bishop, J. (1984) Structure of mouse major urinary protein genes: different splicing configurations in the 3' non-coding region. EMBO J. 3, 1045-1052.

Clark, A. J., Simons, P., Wilmut, I. and Lathe, R. (1987) Pharmaceuticals from transgenic livestock. TIBTECH 5, 20-24.

Cockerill, P. N. and Garrard, W. T. (1987). Chromosomal loop anchorage of the kappa immunuglobulin gene occurs next to the enhancer in a region containing topoisomerase II sites. Cell 44, 273-282.

Conti, A., Liberatori, J., Elia, B. and Cauvin, E. (1977). Isolation of two genetic variants of sheep beta-lactoglobulin by preparative flat-bed isoelectric focusing in granulated gel. Science Tools, The LKB Instrument Journal 24, 54-55.

Cooper, R. and Papaconstantinou, J. (1986) Evidence for the existence of multiple α1-acid glycoprotein genes in the mouse. J. Biol. Chem. 261, 1849-1853.

Corden, J. L., Cadena, D. L., Ahearn (Jr.), J. M. and Dahmus, M. E. (1985). A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. Proc. Natl. Acad. Sci. USA. 82, 7934-7938.

Cowie, A. T. and Tindal, J. S. (1971). In, Physiology of lactation. Publ. Edward Arnold, London.

Cowie, A. T. (1984). In, Reproduction in mammals: 3 Hormonal control of reproduction. Edited by C. R. Austin and R. V. Short. Published by Cambridge University Press, Cambridge.

Craik, C. S., Sprang, S., Fletterick, R. and Rutter, W. J. (1982). Intron-exon splice junctions map at protein surfaces. Nature 299, 180-182.

Craik, C. S., Rutter, W. J. and Fletterick, R. (1983). Splice junctions: Association with variation in protein structure. Science 220, 1125-1129.

Dalgleish, D. G. (1982). Milk proteins - Chemistry and physics. In, Developments in dairy chemistry-1. Ed. P. F. Fox. Publ. Applied Science Publishers Ltd., London.

Dandekar, A. N., Robinson, E. A., Appella, E. and Qasba, P. K. (1982). Complete sequence analysis of cDNA clones encoding rat whey phosphoprotein: homology to a protease inhibitor. Proc. Natl. Acad. Sci. USA 79, 2987-3991.

Daniel, C. W. and Silberstein, G. B. (1987). Postnatal development of the rodent mammary gland. In, The mammary gland: development, regulation and function. Eds, M. C. Neville and Daniel, C. W. Published by Plenum Press, New York.

Davidson, I., Xiao, J. H., Rosales, R., Staub, A. and Chambon, P. (1988). The HeLa cell protein TEF-I binds specifically and cooperatively to two SV40 enhancer motifs of unrelated sequence. Cell 54, 931-942.

Davis, L. G., Dibner, M. D. and Battey, J. F. (1986). In, Basic methods in molecular biology. Elsevier Science Publishing Co. Inc.

Dayhoff, M. O. (1969). Ed. Atlas of protein sequence and structure. National Biomedical Research Foundation. Silver Springs, MD.

Dayhoff, M. O. (1978). Atlas of protein sequence and structure, Vol. 5, suppl. 3, National Biomedical Research Foundation, Washington D.C.

Dayhoff, M. O., Barker, W. C. and Hunt, L. T. (1983). Establishing homologies in protein sequences. Methods in enzymology 91, 524-545.

Dembinski, T. C. and Shiu, R. P. C. (1987). Growth factors in mammary gland

development. In, The mammary gland: development, regulation and function. Eds, M. C. Neville and Daniel, C. W. Published by Plenum Press, New York.

Demmer, L. A., Birkenmeier, E. H., Sweetser, D. A., Levin, M. S., Zollman, S., Sparkes, R. S., Mohandas, T., Lusis, A. J. and Gordon, J. I. (1987) The cellular retinol-binding protein II gene. J. Biol. Chem. 262, 2458-2467.

Denamur, R. (1974). Ribonucleic acids and ribonucleoprotein particles of the mammary gland. In, Lactation, a comprehensive treatise volume I: the mammary gland/development and maintenance. Edited by B. L. Larson and Smith, V. R. Published by Academic Press, New York and London.

Dente, L., Ciliberto, G. and Cortese, R. (1985) Structure of the human α1-acid glycoprotein gene: sequence homology with other human acute phase protein genes. Nucleic Acids Research 13, 3941-3952.

Dente, L., Pizza, M. G., Metspalu, A. and Cortese, R. (1987) Structure and expression of the genes coding for human α1-acid glycoprotein. EMBO J. 6, 2289-2296.

Devereux, J., Haeberli, P. and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Research 12, 387-395.

Devinoy, E., Hubert, C., Jolivet, G., Thepot, D., Clergue, N., Desaleux, M., Dion, M., Servely, J.-L. and Houdebine, L.-M. (1988). Recent data on the structure of rabbit milk protein genes and on the mechanism of the hormonal control of their expression. Reprod. Nutr. Develop. 28, 1145-116.

Dilley, W. G. and Nandi, S. (1968). Rat mammary gland differentiation i*n vitro* in the absence of steroids. Science 161, 59-60.

Dils, R. and Forsyth, I. A. (1981). Preparation and culture of mammary gland explants. Methods in Enzymology 72, 724-742.

Doolittle, R. F. (1981). Similar amino-acid sequences:chance or common ancestry? Science 214, 149-159.

Dorn, A., Bollekens, J., Staub, A., Benoist, C. and Mathis, D. (1987). A multiplicity of CCAAT box-binding proteins. Cell 50, 863-872.

Drayna, D., Fielding, C., McLean, J., Baer, B., Castro, G., Chen, E., Comstock, L., Henzel, W., Kohr, W., Rhee, L., Wion, K. and Lawn, R. (1986). Cloning and expression of human apolipoprotein-D cDNA. J. Biol. Chem. 261, 16535-16539.

Drayna, D. T., McLean, J. W., Wion, K. L., Trent, J., Drabkin, H. A. and Lawn, R. A. (1987). Human apolipoprotein-D gene: Gene sequence, chromosome localisation and homology to the α_{2u}-globulin superfamily. DNA 6, 199-204.

Drenth, J., Low, B. W., Richardson, J. S. and Wright, C. S. (1980). The toxin-agglutinin fold: a new group of small protein structures organised around a four-disulphide core. J. Biol. Chem. 255, 2652-2655.

Dudler, R. and Travers, A. A. (1984). Upstream elements necesary for optimal

function of the *hsp70* promoter in transformed flies. Cell 38, 391-398.

Duncan, C. H. (1987). Novel Alu-type repeat in artiodactyls. Nucleic Acids Research 15, 1340.

Eck, R. V. and Dayhoff, M. O. (1966). Atlas of protein sequence and structure 1966. National Biomedical Research Foundation, Silver Spring, Maryland, USA.

Ehrlich, H. A., Stetler, D., Sheng-Dong, R. Ness, D. and Grumont, C. (1983). Segregation and mapping analysis of polymorphic HLA class I restriction fragments: detection of a novel fragment. Science 222, 72-74.

Emerman, J. T. and Pitelka, D. R. (1977). Maintenance and induction of morphological differentiation in dissociated mammary epithelium on floating collagen membranes. *In Vitro* 13, 316-328.

Emerman, J. M., Enami, J., Pitelka, D. R. and Nandi, S. (1977). Hormonal effect on intracellular and secreted casein in cultures of mouse mammary epithelial cells on floating collagen membranes. Proc. Natl. Acad. Sci. USA 74, 4466-4470.

Eng, C. E. L. and Strom, C. M. (1985). Analysis of three restriction fragment lenngth polymorphisms in the human type II procollagen gene. Amm. J. Hum. Genet. 37, 719-732.

Evans, R. M. (1988). The steroid and thyroid hormone receptor superfamily. Science 240, 889-895.

Feinberg, A. P. and Vogelstein, B. (1983). A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. Analyt. Biochem. 132, 6-13.

Feinberg, A. P. and Vogelstein, B. (1984). Addendum "A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity." Analyt. Biochem. 137, 266-267.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17, 368-376.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783-791.

Felsenstein, J. (1988). Perils of molecular introspection. Nature 335, 118.

Fink, J. S., Verhave, M., Kasper, S., Tsukada, T., Mandel, G. and Goodman, R. H. (1988). The CGTCA sequence motif is essential for biological activity of the vasoactive intestinal peptide gene cAMP-regulated enhancer. Proc. Natl. Acad. Sci. USA 85, 6662-6666.

Fitch, W. M. (1971). Towards defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20, 406-416.

Fitch, W. M and Margoliash, E. (1967). Construction of phylogenetic trees. Science

155, 279-284.

Fitzgerald, M. and Shenk, T. (1981). The sequence 5'-AAUAAA-3' forms part of the recognition site for polyadenylation of late SV40 mRNAs. Cell 24, 251-260.

Fleming, J. R., Head, H. H., Bachman, K. C., Becker, H. N. and Wilcox, C. J. (1986). Induction of lactation: histological and biochemical development of mammary tissue and milk yields of cows injected with estradiol-17β and progesterone for 21 days. J. Dairy Sci. 69, 3008-3021.

Forsyth, I. A. (1983). The endocrinology of lactation. In, Biochemistry of lactation, edited by T. B. Mepham. Elsevier Science Publishers, B. V.

Friedman, M. J. (1983). Control of malaria virulence by α1-acid glycoprotein (orosomucoid), an acute-phase (inflammatory) reactant. Proc. Natl. Acad. Sci. USA. 80, 5421-5424.

Frischauf, A.-M., Lehrach, H., Poustka, A. and Murray, N. (1983). Lambda replacement vectors carrying polylinker sequences. J. Mol. Biol. 170, 827-842.

Fugate, R. D. and Song, P-L. (1980) Spectroscopic characteristics of β-lactoglobulin-retinol complex. Biochim. Biophys. Acta 652, 28-42.

Futterman, S. and Heller, J. (1972) The enhancement of fluorescence and the decreased susceptibility to enzymatic oxidation of retinol complexed with bovine serum albumin, β-lactoglobulin and the retinol-binding protein of human plasma. J.

Biol. Chem. 247, 5168-5172.

Ganguly, M., Carnighan, R. H. and Westphal, U. (1967). Steroid-protein interactions. XIV. Interaction between human α1-acid glycoprotein and progesterone. Biochemistry 6, 2803-2814.

Ganguly, R., Mehta, N. M., Ganguly, N. and Banerjee, M. R. (1979). Glucocorticoid modulation of casein gene transcription in mouse mammary gland. Proc. Natl. Acad. Sci. USA. 76, 6466-6470.

Ganguly, R., Ganguly, N., Mehta, N. M. and Banerjee, M. R. (1980). Absolute requirement of glucocorticoid for expression of the casein gene in the presence of prolactin. Proc. Natl. Acad. Sci. USA. 77, 6003-6006.

Ganguly, R., Majumder, P. K., Ganguly, N. and Banerjee, M. R. (1982). The mechanism of progesterone-glucocorticoid interaction in regulation of casein gene expression. J. Biol. Chem. 257, 2182-2187.

Gasser, S. M. and Laemmli, U. K. (1986). Cohabitation of scaffold binding regions with upstream/enhancer elements of three developmentally regulated genes of *D. melanogaster*. Cell 46, 521-530.

Gasser, S. M. and Laemmli, U. K. (1987). A glimpse at chromosomal order. Trends In Genetics 3, 16-22.

Gaye, P. Viennot, N. and Denamur, R. (1972). *In vitro* synthesis of α-lactalbumin and

β-lactoglobulin by microsome and bound polyribosomes from the mammary gland of lactating sheep. Biochim. Biophys. Acta. 262, 371-380.

Gaye, P., Hue-Delahaie,D., Mercier, J-C., Soulier, S., Vilotte, J-L. and Furet, J-P. (1986). Ovine β-lactoglobulin messenger RNA: Nucleotide sequence and mRNA levels during functional differentiation of the mammary gland. Biochimie 68, 1097-1107.

Gaye, P., Hue-Delahaie, D., Mercier, J.-C., Soulier, S., Vilotte, J.-L. and Furet, J. P. (1987). Complete nucleotide sequence of ovine alpha-lactalbumin mRNA. Biochimie 69, 601-608.

Gaye, P., Mercier, J-C., Petrissant, G., Vilotte, J-L. and Popescu, P. (1986). Structure des ADN complementaires des lactoproteines: application a la recherche des genes et a leur localisation chromosomique. Reprod. Nutr. Develop. 26, 573-574.

Ghazal, P., Clark, A. J. and Bishop, J. O. (1985). Evolutionary amplification of a pseudogene. Proc. Natl. Acad. Sci. USA 82, 4182-4185.

Gilbert, W. (1978). Why genes in pieces? Nature 271, 501.

Gillies, S. D., Morrison, S. L., Oi, V. T. and Tonegawa, S. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. Cell 33, 717-728.

Gilman, M. Z. (1988). The *c-fos* serum response element responds to protein kinase C-dependent and -independent signals but not to cyclic AMP. Genes and Development 2,

394-402.

Glisin, V., Crkvenjakov, R. and Byus, C. (1974). Ribonucleic acid isolation by caesium chloride centrifugation. Biochemistry 13, 2633-2637.

Godovac-Zimmermann, J. (1988). The structural motif of β-lactoglobulin and retinol-binding protein: a basic framework for binding and transport of small hydrophobic molecules? Trends In Biochemical Sciences 13, 64-66.

Godovac-Zimmerman, J., Conti, A., Liberatori, J. and Braunitzer, G. (1985) Homology between the primary structures of β-lactoglobulins and human retinol-binding protein: evidence for a similar biological function? Biol. Chem. Hoppe Seyler 366, 431-434.

Godovac-Zimmerman, J. and Shaw, D. (1987). β-lactoglobulin identified in marsupial milk: the primary structure, binding site and possible function of β-lactoglobulin from Eastern Grey Kangaroo (*macropus giganteus*). Biol. Chem. Hoppe-Seyler 368, 879-886.

Goodburn, S. Zinn, K. and Maniatis, T. (1985) Human β-interferon gene expression is regulated by an inducible enhancer element. Cell 41, 509-520.

Gordon, K., Lee, E., Vitale, J. A., Smith, A. E., Westphal, H. and Hennighausen, L. (1987). Production of human tissue plasminogen activator in transgenic mouse milk. Biotechnology 5, 1183-1187.

Gorodetsky, S. I., Tkach, T. M. and Kapelinskaya, T. V. (1988). Isolation and characterisation of the *Bos taurus* β-casein gene. Gene 66, 87-96.

Green, S. and Chambon, P. (1986). Nuclear recpetors enhance our understanding of transcription regulation. Trends In Genetics 4, 309-314.

Grosveld, F., van Assandelft, G. B., Greaves, D. R. and Kollias, G. (1987). Position-independent, high level expression of the human β-globin gene in transgenic mice. Cell 51, 975-985.

Grosvenor,C. E. and Mena,F. (1974). Neural and hormonal control of milk secretion and milk ejection. In, Lactation, a comprehensive treatise volume I: the mammary gland/development and maintenance. Edited by B. L. Larson and Smith, V. R. Published by Academic Press, New York and London.

Grubb, A., Mendez, E., Fernandez-Luna, J. L., Lopez, C., Mihaesco, E. and Vaerman, J-P. (1986). The molecular organisation of the protein HC-IgA complex (HC-IgA). J. Biol. Chem. 261, 14313-14320.

Gupta, P., Rosen, J. M., D-Eustachio, P. and Ruddle, F. H. (1982). Localization of the casein gene family to a single mouse chromosome. J. Cell Biology 93, 199-204.

Guyette, W. A., Matusik, R. A. and Rosen, J. M. (1979). Prolactin-mediated transcriptional and post-transcriptional control of casein gene expression. Cell 17, 1013-1023.

Haefliger, J-A., Jenne, D., Stanley, K. K. and Tschopp, J. (1987) Structural homology of human complement component c8γ and plasma protein HC: identity of the cysteine bond pattern. Biochim. Biophys. Res. Commun. 149, 750-754.

Hai, T., Liu, F., Allegretto, E. A., Karin, M. and Green, M. R. (1988). A family of immunologically related transcription factors that includes multiple forms of ATF and AP-1. Genes and Development 2, 1216-1226.

Hall, L., Emery, D. C., Davies, M. S., Parker, D. and Craig, R. K. (1987). Organization and sequence of the human α-lactalbumin gene. Biochem. J. 242, 735-742.

Hambraeus, L. (1982) Nutritional aspects of milk proteins. In, Developments in dairy chemistry-1. Ed. P. F. Fox. Publ. Applied Science Publishers Ltd., London.

Hanahan, D. (1983). Studies on transformation of *E. coli* with plasmids. J. Mol. Biol. 166, 557-580.

Harris, S., Ali, S., Anderson, S., Archibald, A. L. and Clark, A. J. (1988). Complete nucleotide sequence of the genomic ovine Beta-lactoglobulin gene. Nucleic Acids Research 16, 10379-10380.

Haslam, S. Z. (1987). Role of sex steroid hormones in normal mammary gland function. In, The mammary gland: development, regulation and function. Eds, M. C. Neville and Daniel, C. W. Published by Plenum Press, New York.

Hatamochi, A., Gloumbek, P. T., Schaftinger, E. V. and Crombrugghe, B. de. (1988). A CCAAT DNA binding factor consisting of two different compenents that are both required for DNA binding. J. Biol. Chem. 263, 5940-5947.

Hayssen, V. and Blackburn, D. G. (1985). α-lactalbumin and the origins of lactation. Evolution 39, 1147-1149.

Heap, R. B. and Flint, A. P. F. (1984). In, Reproduction in mammals: 3 Hormonal control of reproduction. Edited by C. R. Austin and R. V. Short. Published by Cambridge University Press, Cambridge.

Hennighausen, L. G. and Sippel, A. E. (1982). Mouse whey acidic protein is a novel member of the family of 'four-disulfide core' proteins. Nucleic Acids Research 10, 2677- 2684.

Hiba, J., Pozo, E. D., Genazzani, A., Pusterla, E., Lanchranjan, I., Sidiropoulos, D. and Gunti, J. (1977). Hormonal mechanism of milk secretion in the newborn. J. Cell. Endocr. Metabol. 44, 973-976.

Hobbs, A. A., Richards, D. A., Kessler, D. J. and Rosen, J. M. (1982). Complex hormonal regulation of rat casein gene expression. J. Biol. Chem. 257, 3598-3605.

Holden, H. M., Rypniewski, W. R., Law, J. H. and Rayment, I. (1987) The molecular structure of insecticyanin from the tobacco hornworm *Manduca sexta L.* at 2.6 A resolution. EMBO J. 6, 1565-1570.

Holler, M., Westin, G., Jiricny, J. and Schaffner, W. (1988). Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated. Genes & Development 2, 1127-1135.

Hollman, K. H. (1974). Cytology and fine structure of the mammary gland. In, Lactation, a comprehensive treatise volume I: the mammary gland/development and maintenance. Edited by B. L. Larson and Smith, V. R. Published by Academic Press, New York and London.

Horikoshi, M. Carey, M. F., Kakidani,H. And Roeder, R. G. (1988). Mechanism of action of a yeast activator: direct effect of GAL4 derivatives on mammalian TFIID promoter interactions. Cell 54, 665-669.

Houdebine, L-M., Djiane, J., Dusanter-Fourt, I., Martel, P., Kelly, P. A., Devinoy, E. and Servely, J-L. (1985). Hormonal action controlling mammary activity. J. Dairy Sci. 68, 489-500.

Howard, O. M. Z., Rao, G. and Sodetz, J. M. (1987). Complementary DNA and derived amino-acid sequence of the β-subunit of human Complement Protein C8: Identification of a close structural and ancestral relationship to the α subunit and C9. Biochemistry 26, 3565-3570.

Huber, R., Schneider, M., Epp, O., Mayr, I., Messerschmidt, A. and Pflugrath, J. (1987) Crystallization, crystal structure analysis and preliminary molecular model of the Bilin Binding Protein from the insect *Pieris brassicae*. J. Mol. Biol. 195, 423-434.

Huhtala, M. L., Seppala, M., Narvanen, A., Palomaki, P., Julkunen, M. and Bohn, H. (1987) Amino-acid sequence homology between human placental protein 14 and β-lactoglobulins from various species. Endocrinology 120, 2620-2622.

Hunt, L. T., Elzanowski, A. and Barker, W. C. (1987) The homology of complement factor C8 gamma chain and alpha-1-microglobulin. Biochim. Biophys. Res. Commun. 149, 282-288.

Igo-Kemenes, T., Horz, W. and Zachau, H. G. (1982). Chromatin. Ann. Rev. Biochem. 51, 89-121.

Jameson, J. L., Deutsch, P. J., Gallagher, G. D., Jaffe, R. C. and Habener, J. F. (1987). Trans-acting factors interact with a cyclic AMP response element to modulate expression of the human Gonadotropin α gene. Mol. Cell. Biol. 7, 3032-3040.

Jeffreys, A. J. (1979). DNA sequence variants in the $^{G}\gamma$-, $^{A}\gamma$-, δ- and β-globin genes of man. Cell 18, 1-10.

Jeltsch, J. M., Roberts, M., Schatz, C., Garnier, J. M., Brown, A. M. C. and Chambon, P. (1987). Structure of the human oestrogen responsive gene pS2. Nucleic Acids Research 19, 1401-1414.

Jenness, R. (1982). Inter-species comparison of milk proteins. In, Developments in dairy chemistry-1. Ed. P. F. Fox. Publ. Applied Science Publishers Ltd., London.

Johnson, M. L., Levy, J., Supowit, S. C., Yu-Lee, L-Y. and Rosen, J. M. (1983). Tissue- and cell-specific casein gene expression. II. Relationship to site-specific DNA methylation. J. Biol. Chem. 258, 10805-10811.

Jolles, P., Loucheux-Lefebvre, M-H. and Henschen, A. (1978). Structural relatedness of κ-casein and fibrinogen γ-chain. J. Mol. Evol. 11, 271-277.

Jones, N. C., Rigby, P. W. J. and Ziff, E. B. (1988). Trans-acting protein factors and the regulation of eukaryotic transcription: lessons from studies on DNA tumour viruses. Genes and Development 2, 267-281.

Jones, T. A., Bergfors, T., Sedzik, J. and Unge, T. (1988) The three-dimensional structure of P2 myelin protein. EMBO J. 7, 1597-1604.

Jones, W. K., Yu-Lee, Y-L., Clift, S. M., Brown, T. L. and Rosen, J. M. (1985). The rat casein multigene family: fine structure and evolution of the β-casein gene. J. Biol. Chem. 260, 7042-7050.

Julkunen, M., Seppala, M. and Janne, O. A. (1988). Complete amino-acid sequence of human placental protein 14: A progesterone-regulated uterine protein homologous to b-lactoglobulins. Proc. Natl. Acad. Sci. USA 85, 8845-8849.

Kan, Y, W. and Dozy, A. M. (1978). Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. Proc. Natl. Acad. Sci. USA 75, 5631-5635.

Kang, Y. C. and Richardson, T. (1988). Molecular cloning and expression of bovine κ-casein in *Escherichia coli*. J. Dairy Sci. 71, 29-40.

Kastern, W., Bjorck, L. and Akerstrom, B. (1986). Developmental and tissue-specific expression of α1-microglobulin mRNA in the rat. J. Biol. Chem 261, 15070-15074.

Kaumeyer, J. F., Polazzi, J. O. and Kotick, M. P. (1986) The mRNA for a protease inhibitor related to the HI-30 domain of inter-α-trypsin inhibitor also encodes α1-microglobulin (protein HC). Nucleic Acids Research 14, 7839-7850.

Kawamura, K., Satow, H., Ik, L. D., Sakai, S., Takada, S. and Obinata, M. (1987). Modulation of the transferred mouse 26K casein gene in mouse L cells by glucocorticoid hormone. J. Biochem 101, 103-110.

Keller, E. B. and Noon, W. A. (1984). Intron splicing: A conserved internal signal in introns of animal pre-mRNAs. Proc. Natl. Acad. Sci. USA. 81, 7417-7420.

Kerkay, J. and Westphal, U. (1968) Steroid-protein interactions. XIX. Complex formation between α1-acid glycoprotein and steroid hormones. Biochim. Biophys. Acta 170, 324-333.

Kieny, M. P., Lathe, R. and Lecocq, J. P. (1983). New versatile cloning and sequencing vectors based on bacteriophage M13. Gene 26, 91-99.

King, J. W. B. (1969). The distribution of sheep β-lactoglobulins. Animal Production 11, 53-57.

Klein, J. (1986). Natural history of the major histocompatibility complex. Publ. John Wiley & Sons, Inc.

Kleinberg, D. L., Todd, J., Babitsky, G. and Greising, J. (1982). Estradiol inhibits prolactin induced α-lactalbumin production in normal primate mammary tissue in vitro. Endocrinology 110, 279-281.

Klein-Hitpass, L., Ryffel, G. U., Heitlinger, E. and Cato, A. C. B. (1988). A 13 bp palindrome is a functional estrogen responsive element and interacts specifically with estrogen receptor. Nucleic Acids Research 16, 647-663.

Knight, C. H. and Peaker, M. (1982). Development of the mammary gland. Reprod, Fertil. 65, 521-536.

Knight, C. H. (1984). Mammary growth and development: strategies of animals and investigators. In, Zoological Society of London Symposia 51. Physiological strategies in lactation. Edited by M. Peaker, R. G. Vernon and C. H. Knight. Published by Academic Press, London.

Knight, C. H., Maltz, E. and Docherty, A. H. (1986). Milk yield and composition in mice: effects of litter size and lactation number. Comp. Biochem. Physiol 84A, 127-133.

Kolde, H-J. and Braunitzer, G. (1983a). The primary structure of ovine β-lactoglobulin. 1. Isolation of the peptides and sequence. Milchwissenschaft 38,

18-20.

Kolde, H-J. and Braunitzer, G (1983b). The primary structure of ovine β-lactoglobulin. 2. Discussion and gentic aspects. Milchwissenschaft 38, 70-72.

Kollias, G., Hurst, J. deBoer, E. and Grosveld, F. (1987). The human β-globin gene contains a downstream developmental specific enhancer. Nucleic Acids Research 15, 5739-5747.

Kozak, M. (1984). Point mutations close to the AUG initiation codon affect the efficiency of translation of rat preproinsulin *in vivo*. Nature 308, 241-246.

Krauter, K., Leinwand, L., D'Eustachio, P., Ruddle, F. and Darnell, Jr., J. E. (1982). Structural genes of the mouse major urinary protein are on chromosome 4. J. Cell. Biol. 94, 414-417.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 304, 412-417.

Kumar, V. and Chambon, P. (1988). The estrogen receptor binds tightly to its responsive element as a ligand-induced homodimer. Cell 55, 145-156.

Laird, J. E., Jack, L., Hall, L., Boulton, A. P., Parker, D. and Craig, R. K. (1988). Structure and expression of the guinea pig α-lactalbumin gene. Biochem. J. 254, 85-94.

Landschulz, W. H., Johnson, P. F. and McKnight, S. L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. Science 240, 1759-1764.

Larsen, A. and Weintraub, H. (1982). An altered DNA conformation detected by S1 nuclease occurs at specific regions in active chick globin chromatin. Cell 29, 609-622.

Lathe, R., Clark, A. J., Archibald, A. L., Bishop, J. O., Simons, P. and Wilmut, I. (1986). Novel products from livestock. In, exploting new technologies in animal breeding. Edited by C. Smith, J. W. B. King and K. C. McKay. pp 99-102. Oxford University Press.

Lathe, R., Vilotte, J-L. and Clark, A. J. (1987). Plasmid and bacteriophage vectors for excision of intact inserts. Gene 57, 193-201.

Laurent, B. C., Nilsson, M. H. L., Bavik, C. O., Jones, T. A., Sundelin, J. and Peterson, P. A. (1985) Characterisation of the rat retinol-binding protein gene and its comparison to the three-dimensional structure of the protein. J. Biol. Chem. 260, 11476-11480.

Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, R. G. and Maniatis, T. (1978). The isolation and characterisation of linked δ- and β-globin genes from a cloned library of human DNA. Cell 15, 1157-1174.

Lech, K., Anderson, K. and Brent, R. (1988). DNA-bound *fos* proteins activate

transcription in yeast. Cell 52, 179-184.

Lee, K-H., Wells, R. G. and Reed, R. R. (1987) Isolation of an olfactory cDNA: similarity to retinol-binding protein suggests a role in olfaction. Science 235, 1053-1056.

Lee, K-F., DeMayo, F. C., Atie, S. H. and Rosen, J. M. (1988). Tissue specific expression of the rat β-casein gene in transgenic mice. Nucleic Acids Research 16, 1027-1041.

Lewin, B. (1983). In, Genes. Published by John Wiley & Sons, Inc.

Lewin, B. (1980). DNA sequence organisation: non-repetitive and repetitive DNA; inverted and tandem repeats. In, Gene Expression volume II, chapters 18 and 19. Published by John Wiley and Sons, Inc.

Li, M. L., Aggeler, J., Farson, D. A., Hatier, C., Hassell, J. and Bissell, M. J. (1987). Influence of a reconstituted basement membrane and its components on casein gene expression and secretion in mouse mammary epithelial cell. Proc. Natl. Acad. Sci. USA 84, 136-140.

Lindsay, S. and Bird, A. P. (1987). Use of restriction enzymes to detect potential gene sequences in mammalian DNA. Nature 327, 336-338.

Lonnerdal, B. (1985). Biochemistry and physiological function of human milk proteins. Am. J. Clinical Nutrition 42, 1299-1317.

Lopez, C., Grubb, A., Soriano, F. and Mendez, E. (1981) The complete amino-acid sequence of human complex-forming glycoprotein heterogeneous in charge (protein HC). Biochim. Biophs. Res. Commun. 103, 919-925.

Lubon, H. and Hennighausen, L. (1987). Nuclear proteins from lactating mammary glands bind to the promoter of a milk protein gene. Nucleic Acids Research 15, 2103-2121.

Lubon, H. and Hennighausen, L. (1988). Conserved region of the rat α-lactalbumin promoter is a target site for protein binding *in vitro*. Biochem. J. 256, 391-396.

Mandel, J. L. and Chambon, P. (1979). DNA methylation: organ-specific variations in the methylation pattern within and around ovalbumin and other chicken genes. Nucleic Acids Research 7, 2081-2104.

Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982). Molecular cloning, a laboratory manual. Published by Cold Spring Harbor Laboratory.

Maniatis, T., Goodburn, S. and Fischer, J. A. (1987). Regulation of inducible and tissue-specific gene expression. Science 236, 1237-1245.

Martinez, E., Givel, F. and Wahli, W. (1987). The estrogen-responsive element as an inducible enhancer: DNA sequence requirements and conversion to a glucocorticoid-responsive element. EMBO J. 6, 3719-3727.

Mathis, D. J. and Chambon, P. (1981). The SV40 early region TATA box is required for accurate in vitro initiation of transcription. Nature 290, 310-315.

Matyukov, V. S. and Urnyshev, A. P. (1980). Linkage between cattle milk α_{s1}-, β–, κ-casein loci. Genetika 16, 884-886.

McKusick, V. A. (1986). The human gene map. Symposia on quantitative biology 51, 1123-1208.

McLauchlan, J., Gaffney, D., Whitton, J. L. and Clements, J. B. (1985). The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. Nucleic Acids Research 13, 1347-1368.

McNeilly, A. S. and Andrews, P. (1974). Purification and characterisation of caprine prolactin. J. Endocrinol. 60, 359-367.

McNeilly, A. S. (1984). Changes in FSH and the pulsatile secretion of LH during the delay in oestrus induced by treatment of ewes with bovine follicular fluid. J. Reprod. Fertil. 72, 165-172.

Mehta, N. M., Ganguly, N., Ganguly, R. and Banerjee, M. H. (1980). Hormonal modulation of the casein gene expression in a mammogenesis-lactogenesis culture model of the whole mammary gland of the mouse. J. Biol. Chem. 255, 4430-4434.

Mendez, E., Fernandez-Luna, J. L., Grubb, A. and Leyva-Cobian, F. (1986). Human protein HC and its IgA complex are inhibitors of neutrophil chemotaxis. Proc. Natl.

Acad. Sci., USA. 83, 1472-1475.

Menon, R. S. and Ham, R. G. (1988). Molecular cloning and apparent polymorphism of human beta-casein cDNA (Abstract). J. Cell Biol. 107, 523a.

Mepham, T. B. (1987). Control of mammary function at the cellular level. In, Physiology of Lactation. Open University Press, Milton Keynes.

Mepham, T. B., Gaye, P. and Mercier, J-C. (1982) In, Developments in dairy chemistry-1. Ed. P. F. Fox. Publ. Applied Science Publishers Ltd., London.

Mercier, J.-C., Haze, G., Gaye, P. and Hue-Delahaie, D. (1978). Amino terminal sequence of the presucrsor of ovine β-lactoglobulin. Biochem. Biophys. Res. Commun. 82, 1236-1245.

Mercier, J.-C. and Gaye, P. (1983). Milk protein synthesis. In, Biochemistry of lactation, edited by T. B. Mepham. Elsevier Science Publishers, B. V.

Mercier, J-C., Gaye, P., Soulier, S., Hue-Delahaie, D. and Vilotte, J-L. (1985). Construction and identification of recombinant plasmids carrying cDNAs coding for ovine α_{s1}-, α_{s2}-, β-, κ-casein and β-lactoglobulin. Nucleotide sequence of α_{s1}-casein cDNA. Biochimie 67, 959-971.

Montell, C., Fisher, E. F., Caruthers, M. H. and Berk, A. J. (1983). Inhibition of RNA cleavage but not polyadenylation by a point mutation in mRNA 3' consensus sequence AAUAAA. Nature 305, 600-605.

Morton, N. E. and Bruns, G. A. (1987). Report of the commitee on the gentic constitution of chromosome 1 and 2. Human Gene Mapping 9 (1987); Ninth International Workshop on Human Gene Mapping, Cytogenet. Cell Genet. 46, 102-130.

Mouchiroud, D., Gautier, C. and Bernardi, G. (1988). The compositional distribution of coding sequences and DNA molecules in humans and murids. J. Mol. Evol. 27, 311-320.

Nahon, J.-L., Venetianer, A. and Sala-Trepat, J. M. (1987). Specific sets of DNaseI-hypersensitive sites are associated with the potential and overt expression of the rat albumin and α-fetoprotein genes. Proc. Natl. Acad. Sci. USA 84, 2135-2139.

Nakhasi, H. L. and Qasba, P. K. (1979). Quantitation of milk proteins and their mRNAs in rat mammary gland at various stages of gestation and lactation. J. Biol. Chem. 254, 6016-6025.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino-acid sequence of two proteins. J. Mol. Biol. 48, 443-453.

Newcomer, M. E., Jones, T. A., Aqvist, J., Sundelin, J., Eriksson. U., Rask, L. and Peterson, P. A. (1984). The three-dimensional structure of retinol-binding protein. EMBO J. 3, 1451-1454.

Ng, S. C., Rao, G., Howard, O. M. and Sodetz, J. M. (1987). The eighth component of human Complement: Evidence that it is an oligomeric serum protein assembled from

products of three different genes. Biochemistry 26, 5229-5233.

Nowock, J., Borgmeyer, U., Puschel, A. W., Rupp, R. A. and Sippel, A. E. (1985). The TGGCA protein binds to the MMTV-LTR, the adenovirus origin of replication and the BK virus enhancer. Nucleic Acids Research 13, 2045-2061.

Nussinov, R. (1986). Sequence signals which may be required for efficient formation of mRNA 3' termini. Nucleic Acids Research 14, 3557-3571.

O'Brien, S. J. (Ed.). Genetic Maps 1984. Publ. by Cold Spring Harbor Laboratory.

Ono, M. and Oka, T. (1980a). The differential actions of cortisol on the accumulation of α-lactalbumin and casein in midpregnant mouse mammary gland in culture. Cell 19, 473-480.

Ono, M. and Oka, T. (1980b). α-lactalbumin-casein induction in virgin mouse mammary explants: dose-dependent differential action of cortisol. Science 207, 1367-1369.

Orkin, S. H. and Kazazian Jr., H. H. (1984). The mutation and polymorphism of the human β-globin gene and its surrounding DNA. Ann. Rev. Genetics 18, 131-171.

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S. and Sharp, P. A. (1986). Splicing of messenger RNA precursors. Ann. Rev. Biochem. 55, 1119-1150.

Palmiter, R. D. and Brinster, R. L. (1986). Germline transformation of mice. Annu.

Rev. Genet. 20, 465-499.

Pan, J., Elder, J. T., Duncan, C. H. and Weissman, S. M. (1981). Structural analysis of interspersed repetitive polymerase III transcription units in human DNA. Nucleic Acids Research 9, 1151-1170.

Papiz, M. Z., Sawyer, L., Eliopoulos, E. E., North, A. C. T., Findlay, J. B. C., Sivaprasadarao, R., Jones, T. A., Newcomer, M. E. and Kraulis, P. J. (1986). The structure of β-lactoglobulin and its similarity to plasma retinol-binding protein. Nature 324, 383-385.

Pelham, H. (1985). Activation of heat shock genes in eukaryotes. Trends In Genetics 1, 31-35.

Pervaiz, S. and Brew, K. (1985) Homology of β-lactoglobulin, serum retinol-binding protein and protein HC. Science 228, 335-337.

Pervaiz, S. and Brew, K. (1987). Homology and structure-function correlations between α1-acid glycoprotein and serum retinol-binding protein and its relatives. FASEB J. 1, 209-214.

Peterlin, B. M., Hardy, K. J. and Larsen, A. S. (1987). Chromatin structure of the HLA-DRα gene in different functional states of major histocompatibility complex class II gene expression. Mol. Cell. Biol. 7, 1967-1972.

Pevsner, J., Sklar, P. B. and Snyder, S. H. (1986). Odorant-binding protein:

localisation to nasal glands and secretions. Proc. Natl. Acad. Sci., USA 83, 4942-4946.

Pevsner, J., Hwang, P. M., Sklar, P. B., Venable, J. C. and Snyder, S. H. (1988a) Odorant-binding protein and its mRNA are localised to lateral nasal gland implying a carrier function. Proc. Natl. Acad. Sci. USA 85, 2383-2387.

Pevsner, J., Randall, R. R., Feinstein, P. G. and Snyder, S. H. (1988b) Molecular cloning of a ligand carrier family. Science 241, 336-339.

Phillips, J. A., Hjelle, B. L., Seeburg, P. H. and Zachmann, M. (1983). Molecular basis for familial isolated growth hormone deficiency type I. Proc. Natl. Acad. Sci. USA 78, 6372-6375.

Pittius, C. W., Sankaran, L., Topper, Y. J. and Hennighausen, L. (1988). Comparison of the regulation of the whey acidic protein gene with that of a hybrid gene containing the whey acidic protein gene promoter in transgenic mice. Mol. Endocrinol. 2, 1027-1032.

Ploegh, H. L., Orr, H. T. and Strominger, J. L. (1980). Molecular cloning of a human histocompatibility antigen cDNA fragment. Proc. Natl. Acad. Sci. USA 77, 6081-6085.

Ponte, P., Gunning, P., BLau, H. and Kedes, L. (1983). Human actin genes are single copy for a-skeletal and a-cardiac actin but multicopf for b- and g-cytoskeletal genes: 3' unstranslated regions are isotype specific but are conserved in evolution. Mol. Cell. Biol. 3, 1783-1791.

Powell, B. C., Sleigh, M. J., Ward, K. A. and Rogers, G. E. (1983). Mammalian keratin gene families: organisation of genes coding for the B2 high-sulphur proteins of sheep wool. Nucleic Acids Research 11, 5327-5346.

Prager, E. M. and Wilson, A. C. (1988). Ancient origin of α-lactalbumin from lysozyme: analysis of DNA and amino-acid sequences. J. Mol. Evol. 27, 326-335.

Prochownik, E. V., Antonarkis, S., Bauer, K. A., Rosenberg, R. D., Fearon, E. R. and Orkin, S. H. (1983). Molecular heterogeneity of inherited antithrombin III deficiency. N. Engl. J. Med. 308, 1549-1552.

Pratt, W. B., Jolly, D. J., Pratt, D. V., Hollenberg, S. M., Giguere, V., Cadepond, F. M., Schweizer-Groyer, G., Catelli, M-G., Evans, R. M. and Baulieu, E-E. (1988). A region in the steroid binding domain determines formation of the non-DNA-binding, 9 S glucocorticoid receptor complex. J. Biol. Chem. 263, 267-273.

Proudfoot, N. J. and Brownlee, G. G. (1976). 3' non-coding region sequences in eukaryotic messenger RNA. Nature 263, 211-214.

Qasba, P. K. and Safaya, S. K. (1984). Similarity of the nucleotide sequences of rat α-lactalbumin and chicken lysozyme genes. Nature 308, 377-380.

Rao, A. G., Howard, O. M. Z., Ng, S. C., Whitehead, A. S., Colten, H. R. and Sodetz, J. M. (1987). Complementary DNA and derived amino-acid sequence of the α subunit of human Complement Protein C8: Evidence for the existence of a separate α subunit messenger RNA. Biochemistry 26, 3556-3564.

Rask, L., Anundi, H., Bohme, J., Eriksson, U., Ronne, H., Sege, K. and Peterson ,P. A. (1981). Structural and functional studies of vitamin A-binding proteins. Annals of the New York Academy of Sciences 359, 79-90.

Razin, A. and Riggs, A. D. (1980). DNA methylation and gene function. Science 210, 604-610.

Razin, A. and Cedar, H. (1984). DNA methylation in eukaryotic cells. Int. Rev. Cytol. 92, 159-185.

Ricca G. A., Hamilton, R. W., McLean, J., Conn, A., Kalinyak, J. E. and Taylor, J. M. (1981). Rat α1-acid glycoprotein mRNA: cloning of double-stranded cDNA and kinetics of induction of mRNA levels following acute inflammation. J. Biol. Chem. 256, 10362-10368.

Rich, A., Nordheim, A. and Wang, A. H.-J. (1984). The chemistry and biochemistry of left-handed Z-DNA. Ann. Rev. Biochem. 53, 791-846.

Riley, C. T., Barbeau, B. K., Keini, P.S., Kezdy, F. J., Heinrikson, R. L. and Law, J. H. (1984) The covalent protein structure of insecticyanin, a blue biliprotein from the hemolymph of the Tobacco Hornworm, *Manduca sexta L.* J. Biol. Chem. 259, 13159-13165.

Rio, M. C., Bellocq, J. P., Daniel, J. Y., Tomasetto, C., Lathe, R., Chenard, M. P., Batzenschlager, A. and Chambon, P. (1988). Breast cancer-associated pS2 protein:

synthesis and secretion by normal stomach mucosa. Science 241, 705-708.

Rocchi, M., Colantuoni, V. and Romeo, G. (1987). Assignment of RBP to 10 and subregional mapping of CRBP on 3q21-3qter (Abstract). Cytogenet. Cell. Genet. 46, 683.

Rogde, S., Olaisen, B., Gedde-Dahl, Jr., T. and Teisberg, P. (1986). The C8A and C8B loci are closely linked on chromosome 1. Ann Hum. Genet. 50, 139-144.

Rogers, J. (1985). Exon shuffling and intron insertion in serine protease genes. Nature 315, 458-459.

Rogers, J. (1986). Introns between protein domains: selective insertion or frameshifting? Trends In Genetics , 223.

Rosen, J. M. (1987). Milk protein gene structure and expression. In, The mammary gland: development, regulation and function. Eds, M. C. Neville and Daniel, C. W. Published by Plenum Press, New York.

Rosen, J. M., Woo, S. L. C. and Comstock, J. P. (1975). Regulation of casein messenger RNA during the development of the rat mammary gland. Biochemistry 14, 2895-2903.

Rosen, J. M., O'Neal, D. L., McHugh, J. E. and Comstock, J. P. (1978). Progesterone-mediated inhibition of casein mRNA and polysomal casein synthesis in the rat mammary gland during pregnancy. Biochemistry 17, 290-297.

Rosen, J. M., Jones, W. K., Campbell, S. M., Bisbee, C. A. and Yu-Lee, L-Y. (1985). Structure and regulation of peptide hormone-responsive genes. Membrane receptors and cellular regulation, pp. 385-396.

Russo, J. and Russo, I. H. (1987). Development of the human mammary gland. In, The mammary gland: development, regulation and function. Eds, M. C. Neville and Daniel, C. W. Published by Plenum Press, New York.

Sacchetini, J. C., Said, B., Schulz, H. and Gordon, J. I. (1986). Rat heart fatty acid-binding protein is highly homologous to the murine adipocyte 422 protein and the P2 protein of peripheral nerve myelin. J. Biol. Chem. 261, 8218-8223.

Sakakura, T., Nishizuka, Y. and Dawe, C. J. (1976). Mesenchyme-dependent morphogenesis and epithelium-specific cytodifferentiation in mouse mammary gland. Science 194, 1439-1441.

Sakakura, T., Sakagani, Y. and Nishizuka, Y. (1982). Dual origin of mesenchymal tissues participating in mouse mammary gland embryogenessis. Developmental Biology 91, 202-207.

Sakakura, T. (1987). Mammary embryogenesis. In, The mammary gland: development, regulation and function. Eds, M. C. Neville and Daniel, C. W. Published by Plenum Press, New York.

Sanger, F., Nicklen, S. and Coulson, A. R. (1978). DNA sequencing with

chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463-5467.

Santoro, C., Mermod, N., Andrews, P. C. and Tjian, R. (1988). A family of human CCAAT-box-binding proteins active in transcription and DNA replication: cloning and expression of multiple cDNAs. Nature 334, 218-224.

Sawyer, L. (1987). One fold among many. Nature 327, 659.

Scheidereit, C., Westphal, H. M., Carlson, C., Bosshard, H. and Beato, M. (1986). Molecular model of the interaction between the glucocorticoid receptor and the regulatory elements of inducible genes. DNA 5, 383-391.

Schmid, K. Burgi, W., Collins, J. H. and Nanno, S. (1974). The disulphide bonds of α1-acid glycoprotein. Biochemistry 13, 2694-2697.

Schmidt, D. G. (1982). Association of caseins and casein micelle structure. In, Developments in dairy chemistry-1. Ed. P. F. Fox. Publ. Applied Science Publishers Ltd., London.

Schonenberger, C.-A., Andres, A.-C., Groner, B., Van der VAlk, M., LeMeur, M. and Gerlinger, P. (1988). Targeted c-myc gene expression in mammary glands of transgenic mice induces mammary tumours with constitutive milk protein gene transcription. EMBO J. 7, 169-175.

Searle, A. G., Peters, J., Lyon, M. F., Evans, E. P., Edwards, J. H. and Buckle, V. J. (1987). Chromosome maps of man and mouse, III. Genomics 1, 3-18.

Serfling, E. Jasin, M. and Schaffner, W. (1985) Enhancers and eukaryotic gene transcription. Trends In Genetics 1, 224-230.

Shahan, K., Denaro, M., Gilmartin, M., Shi, Y. and Derman, E. (1987). Expression of six mouse major urinary protein genes in the mammary, parotid, sublingual, submaxillary, and lachrymal glands and in the liver. Mol Cell. Biol. 7, 1947-1954.

Shaw, P. H., Held, W. A. and Hastie, N. D. (1983). The gene family for Major Urinary Proteins: Expression in several secretory tissues of the mouse. Cell 32, 755-761.

Shewale, J. G., Sinha, S. K. and Brew, K. (1984). Evolution of α-lactalbumins. J. Biol. Chem. 259, 4947-4956.

Shuster, R. C., Houdebine, L-M. and Gaye, P. (1976). Studies on the synthesis of casein messenger RNA during pregnancy in the rabbit. Eur. J. Biochem. 71, 193-199.

Sigler, P. B. (1988). Acid blobs and negative noodles. Nature 333, 210-212.

Simons, J. P., McClenaghan, M. and Clark, A. J. (1987). Alteration of the quality of milk by expression of sheep β-lactoglobulin in transgenic mice. Nature 328, 530-532.

Simons, J. P., Wilmut, I., Clark, A. J., Archibald, A. L., Bishop, J. O. and Lathe, R. (1988). Gene transfer into sheep. Biotechnology 6, 179-183.

Singer M. F. (1981). Highly repeated sequences in mammalian genomes. International Review of Cytology 76, 67-112.

Sinha, Y. N. and Tucker, H. A. (1969a). Relationship of pituitary prolactin and LH to mammary and uterine growth of pubertal rats during the oestrus cycle. Proc. Soc. Exp. Biol. Med. 131, 908-913.

Sinha, Y. N. and Tucker, H. A. (1969b). mammary development and pituitary prolactin level of heifers from birth through puberty and during the oestrus cycle. J. Dairy Sci. 52, 507-512.

Skowronski, J., Pluciennicak, A., Bednarek, A. and Jaworski, J. (1984). Bovine 1.709 Satellite recombination hotspots and dispersed repeated sequences. J. Mol. Biol. 177, 399-416.

Slightom, J. L., Blechl, A. E. and Smithies, O. (1980). Human foetal $^{G}\gamma$- and $^{A}\gamma$-globin genes: complete nucleotide sequencesm suggest that DNA can be exchanged between duplicated genes. Cell 21, 627-638.

Smith, S. G., Lewis, M., Aschaffenburg, R., Fenna, R. G., Wilson, I. A., Sundralingam, M., Stuart, D. I. and Phillips, D. C. (1987). Crystallographic analysis of the three-dimensional structure of baboon α-lactalbumin at low resolution. Homology with lysozyme. Biochem. J. 242, 353-360.

Sollner-Webb, B. and Reeder, R. H. (1979). The nucleotide sequences of the initiation and termination sites for ribosomal RNA transcription in *X. laevis.* Cell 18, 485-499.

Sorger, P. K. and Pelham, H. R. B. (1988). Yeast heat shock factor is an essential DNA-binding protein that exhibits temperature-dependent phosphorylation. Cell 54, 855-864.

Spence, S. E., Young, R. M., Garner, K. J. and Lingrel, J. B. (1985). Localisation and characterisation of members of a family of repetitive sequences in the goat beta-globin locus. Nucleic Acids Research 13, 2171-2186.

Staden, R. (1982). An interactive program for comparing and aligning nucleic acid and amino-acid sequences. Nucleic Acids Research 10, 2951-2961.

Stalder, J., Groudine, M., Dodgson, J. B., Engel, J. D. and Weintraub, H. (1980). Hb switching in chickens. Cell 19, 973-980.

Steiner., D. F., Chan, S. J., Welsh, J. M. and Kwok, S. C. M. (1985). Structure and evolution of the insulin gene. Annual Review of Genetics 19, 463-484.

Stewart, A. F., Willis, I. M. and MacKinlay, A. G. (1984). Nucleotide sequences of bovine α_{s1}- and κ-casein cDNAs. Nucleic Acids Research 12, 3895-3907.

Stewart, A. F., Bonsing, J., Beattie, C. W., Shah, F., Willis, I. M. and MacKinlay, A. G. (1987). Complete nucleotide sequences of bovine α_{s2}- and β-casein cDNAs: comparisons with related sequences in other species. Mol. Biol. Evol. 4, 231-241.

Strahle, U., Klock, G. and Schutz, G. (1987) A DNA sequence of 15 base pairs is

sufficient to mediate both glucocorticoid and progesterone induction of gene expression. Proc. Natl. Acad. Sci. USA 84, 7871-7875.

Struhl, K. (1986). Constitutive and inducible Saccharomyces cerevisiae promoters: evidence for two distinct molecular mechanisms. Mol. Cell. Biol. 6, 3847-3853.

Stuart, G. W., Searle, P. F. and Palmiter, R. D. (1985). Identification of multiple metal regulatory elements in mouse metallothionein-I promoter by assaying synthetic sequences. Nature 317, 828-831.

Suard, Y. M. L., Haeuptle, M.-T, Farinon, E. and Kraehenbuhl, J. P. (1983). Cell proliferation and milk protein gene expression in rabbit mammary cell cultures. J. Cell Biol. 96, 1435-1442.

Swaisgood, H. E. (1982). Chemistry of milk protein. In, Developments in dairy chemistry-1. Ed. P. F. Fox. Publ. Applied Science Publishers Ltd., London.

Takahashi, K., Odani, S. Ono, T. (1982). A close structural relationship of rat liver z-protein to cellular retinoid-binding proteins and peripheral nerve myelin P2 protein. Biochim Biophys. Res. Commun. 106,1099-1105.

Tejler, L. and Grubb, A. O. (1976). A complex-forming glycoprotein heterogeneous in charge and present in human plasma, urine and cerebrospinal fluid. Biochem. Biophys. Acta 439, 82-94.

Teyssot, B. and Houdebine, L-M. (1980). Role of prolactin in the transcription of

β-casein and 28S ribosomal genes in the rabbit mammary gland. Eur. J. Biochem. 110, 263-272.

Teyssot, B. and Houdebine, L-M. (1981). Role of progesterone and glucocorticoids in the transcription of β-casein and 28S ribosomal genes in the rabbit mammary gland. Eur. J. Biochem. 114, 597-608.

Thompson, M. D., Dave, J. R. and Nakhasi, H. L. (1985). Molecular cloning of mouse mammary gland κ-casein: comparison with rat κ-casein and rat and human γ-fibrinogen. DNA 4, 263-271.

Thompson, M. P. and Farrell, H. M., Jr. (1974). In, Lactation III. Eds. B. L. Larson and V. R. Smith. Publ. Academic Press, New York.

Thordarson, G. and Talamantes, F. (1987). Role of the placenta in mammary gland development and function. In, The mammary gland: development, regulation and function. Eds, M. C. Neville and Daniel, C. W. Published by Plenum Press, New York.

Topper, Y. J. and Freeman, C. S. (1980). Multiple hormone interactions in the developmental biology of the mammary gland. Physiol. Reviews 60, 1049-1106.

Traboni, C. and Cortese, R. (1986) Sequence of a full length cDNA coding for human protein HC (α1-microglobulin). Nucleic Acids Research 14, 6340.

Traboni, C., Tosini, F., Romeo, G. and Rocchi, M. (1987). The gene coding for protein HC and HI-30 domain of inter-α-trypsin inhibitor maps on chromosome 9

(Abstract). Cytogenet. Cell. Genet. 46, 705.

Tsai, S. Y., Carlstedt-Duke, J., Weigel, N. L., Dahlman, K., Gustaffson, J.-A., Tsai, M.-J. and O-Malley, B. W. (1988). Molecular interactions of steroid hormone receptor with its enhancer element: evidence for receptor dimer formation. Cell 55, 361-369.

Tucker, H. A. (1974). General endocrinological control of lactation. In, Lactation, a comprehensive treatise volume I: the mammary gland/development and maintenance. Edited by B. L. Larson and Smith, V. R. Published by Academic Press, New York and London.

Turcotte, B., Guertin, M., Chevrette, M., LaRue, H. and Belanger, L. (1986). DNaseI hypersensitivity and methylation of the 5' flanking regions of the a-fetoprotein gene during development and glucocorticoid-induced repression of its activity in rat liver. Nucleic Acids Research 14, 9827-9841.

Unterman, R. D., Lynch, K. R., Nakhasi, H. L., Dolan, K. P., Hamilton, J. W., Cohn, D. V. and Feigelson, P. (1981). Cloning and sequence of several α_{2u}-globulin cDNAs. Proc. Natl. Acad. Sci. USA. 78, 3478-3482.

Vaiman, M., Chardon, P. and Cohen, D. (1986). DNA polymorphism in the major histocompatibility complex of man and various farm animals. Animal Genetics 17, 113-133.

Vanaman, T. C., Brew, K. and Hill, R. L. (1970) The disulfide bonds of bovine

α-lactalbumin. J. Biol. Chem. 245, 4583-4590.

Vandenbergh, J. G., Whitsett, J. M. and Lombardi, J. R. (1975). Partial isolation of a pheromone accelerating puberty in female mice. J. Reprod. Fertil. 43, 515-523.

van der Ploeg, L. H. T. and Flavell. R. A. (1980). DNA methylation in the γ-globin locus in erythroid and nonerythroid tissues. Cell 19, 947-958.

van Tienhoven, A. (1968). Reproductive physiology of vertabrates. Publ. W. B. Saunders Company, Philadelphia.

Vilotte, J-L., Soulier, S., Mercier, J-C., Gaye, P., Hue-Delahaie, D. and Furet, J-P. (1987). Complete nucleotide sequence of bovine α-lactalbumin gene. Comparison with its rat counterpart. Biochimie 69, 609-620.

Vogt, P. K., Bos, T. J. and Doolittle, R. F. (1987). Homology between the DNA-binding domain of the GCN4 regulatory protein of yeast and the carboxyl-terminal region of a protein coded for by the oncogene *jun*. Proc. Natl. Acad. Sci. USA. 84, 3316-3319.

von der Ahe, D. Janich, S., Scheidereit, C., Renkawitz, R., Schutz, G. and Beato, M. (1985) Glucocorticoid and progesterone receptors bind to the same sites in two hormonally regulated promoters. Nature 313, 706-709.

von der Ahe, D., Renoir, J. M., Bouchou, T., Baulieu, E-E. and Beato, M. (1986) Receptors for glucocorticosteroid and progesterone recognise distinct features of a DNA regulatory element. Proc. Natl. Acad. Sci. USA 83, 2817-2821.

Vonderhaar, B. K. and Nakhasi, H. L. (1986). Bifunctional activity of EGF on α- and κ-casein gene expression in rodent mammary in vitro. Endocrinology 119, 1178-1184.

Vonderhaar, B. K. (1987). Prolactin: transport, function and receptors in mammary gland development and differentiation. In, The mammary gland: development, regulation and function. Eds, M. C. Neville and Daniel, C. W. Published by Plenum Press, New York.

Wagh, P. V., Bornstein, I. and Winzler, R. J. (1969). The structure of a glycopeptide from human orosomucoid (α_1-acid glycoprotein). J. Biol. Chem. 244, 658-665.

Walstra, P. and Jenness, R. (1984) Dairy Chemistry and Physics. Publ. John Wiley & Sons, Inc.

Watanabe, Y., Tsukado, T., Notake, M., Nakanishi, S. and Numa, S. (1982). Structural analysis of repetitive DNA sequences in the bovine corticotropin-β-lipotropin precursor gene region. Nucleic Acids Research 10, 1459-1469.

Webb, G. C., Earle, M. E., Merritt, C. and Board, P. G. (1988). Localisation of human α1-acid glycoprotein genes to 9q31-q34.1. Cytogenet. Cell Genet. 47, 18-21.

Weech, P. K., Camato, R., Milne, R. W. and Marcel, Y. L. (1986) Apolipoprotein D and cross-reacting human plasma apolipoproteins identified using monoclonal antibodies. J. Biol. Chem. 261, 7941-7951.

Wistow, G. and Piatigorsky, J. (1988). Lens crystallins: the evolution and expression of proteins for a highly specialised tissue. Ann. Rev. Biochem. 57, 479-504.

Woo, S. L. C., Lidsky, A. S., Guttler, F., Thirumalachary, C., Robson and K. J. H. (1984). Prenatal diagnosis of classical phenylketonuria by gene mapping. J. Am. Med. Assoc. 25, 1998-2002.

Wood, B. G., Washburn, L. L., Makherjee, A. S. and Banerjee, M. R. (1975). Hormonal regulation of lobulo-alveolar growth, functional differentiation and regreeion of whole mammary gland in organ culture. J. Endocr. 65, 1-6.

Wu, C. (1980). The 5' ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNaseI. Nature 286, 854-860.

Yang, J., Richards, J., Bowman, P., Guzman, R., Enami, J., McCormick, K., Hamamoto, S., Pitelka, D. and Nandi, S. (1979). Sustained growth and three-dimensional organisation of primary mammary tumour epithelial cells embedded in collagen gels. Proc. Natl. Acad. Sci. USA 76, 3401-3405.

Yanisch-Perron, C., Vieira, J. and Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13 mp18 and pUC19 vectors. Gene 33, 103-119.

Yu-Lee, L-Y. and Rosen, J. M. (1983). The rat casein multigene family I. Fine structure of the γ-casein gene. J. Biol. Chem. 258, 10794-10804.

Yu-Lee, L-Y., Richter-Mann, L., Couch, C. H., Stewart, A. F., Mackinlay, A. G. and Rosen, J. M. (1986). Evolution of the casein multigene family: conserved sequences in the 5' flanking and exon regions. Nucleic Acids Research 14, 1883-1902.

Zheng, X.-M., Moncollin, V., Egly, J.-M. and Chambon, P. (1987). A general transcription factor forms a stable complex with RNA pol B (II). Cell 50, 361-368.

Reprinted from *J. Mol. Biol.* (1988) **199**, 415–426

Characterization of the Gene Encoding Ovine Beta-lactoglobulin

Similarity to the Genes for Retinol Binding Protein and Other Secretory Proteins

Simak Ali and A. John Clark

J. Mol. Biol. (1988) **199**, 415–426

Characterization of the Gene Encoding Ovine Beta-lactoglobulin

Similarity to the Genes for Retinol Binding Protein and Other Secretory Proteins

Simak Ali and A. John Clark

AFRC Institute of Animal Physiology and Genetics Research
Kings Buildings, West Mains Road
Edinburgh EH9 3JQ, Scotland

(Received 25 June 1987, and in revised form 23 September 1987)

Beta-lactoglobulin is the major whey protein in the milk of ruminants and is expressed in the mammary gland during pregnancy and lactation. Here we describe the isolation and characterization of genomic clones encoding ovine beta-lactoglobulin. Two very similar but non-identical, types of beta-lactoglobulin clone were obtained. DNA sequence analysis of one of these showed that the gene is 4900 bases long and contains seven exons. It codes for a protein of 180 amino acid residues, containing an 18-residue signal peptide, within exons I to VI; exon VII is non-coding. We show that the genes encoding serum retinol binding protein, major urinary protein, alpha-1-acid glycoprotein and apolipoprotein D have a similar organization of exons and introns to beta-lactoglobulin. In particular, a comparison between beta-lactoglobulin and retinol binding protein shows that both genes encode equivalent elements of three-dimensional protein structure within analogous exons. These proteins are all members of a large, diverse family of secretory proteins, many of which function in binding small hydrophobic molecules.

1. Introduction

Beta-lactoglobulin (BLG†) is the major component of the milk whey of ruminants and is found in the milk of other animals, including horses, pigs, dogs and dolphins (Jenness, 1982; Pervaiz & Brew, 1985). In ruminants, BLG consists of a mature polypeptide chain of 162 amino acid residues containing five cysteine residues, four of which are involved in forming intra-chain disulphide bridges. In the milk of ruminants, BLG exists predominantly as a stable dimer. It binds a variety of hydrophobic molecules (Lovrien & Anderson, 1969; Futterman & Heller, 1972), including retinol (Cogan *et al.*, 1976; Fugate & Song, 1980), and it has been proposed that BLG may function in the transport of this molecule (Papiz *et al.*, 1986).

The three-dimensional structure of bovine BLG has been determined (Papiz *et al.*, 1986) at a resolution of 2·8 Å (1 Å = 0·1 nm). The most striking feature of the molecule is a β-barrel core composed of two slabs of antiparallel β-sheet, which shows a remarkable similarity to the structure of human serum retinol binding protein (RBP: Newcomer *et al.*, 1984). When the two molecules are superimposed, 129 out of 162 residues can be aligned closely. Despite the close similarity of their three-dimensional structures, BLG and RBP share only limited amino acid sequence homology (Godovac-Zimmerman *et al.*, 1985).

The gene encoding BLG is thought to be single copy in ruminants and is expressed in the mammary gland during pregnancy and lactation (Aschaffenburg & Drewry, 1957; Gaye *et al.*, 1986). Ovine BLG transcripts are about 800 nucleotides long and contain approximately 5 % poly(A)$^+$ RNA (Mercier *et al.*, 1985).

We describe the isolation and characterization of a functional gene encoding ovine BLG. Its similarity with the gene encoding rat RBP is described, with particular reference to the relationship between gene organization and three-

† Abbreviations used: BLG, beta-lactoglobulin; RBP, retinol binding protein; kb, 10^3 bases or base-pairs; bp, base-pair(s); α2-UG, alpha-2-urinary globulin; MUP, major urinary protein; apo-D, apolipoprotein D; α1-AGP, alpha-1-acid glycoprotein; HCHU, human protein HC; BG, protein from frog Bowman's gland; INCYN, tobacco hornworm insecticyanin; ESP, rat epididymal secretory protein.

0022–2836/88/030415–12 $03.00/0

dimensional protein structure. We show that rodent urinary protein, alpha-1-acid glycoprotein, and apolipoprotein D genes have a similar arrangement of exons and introns, and suggest that the genes encoding a variety of other distantly related secretory proteins may be organized similarly.

2. Materials and Methods

(a) *Library construction*

High molecular weight DNA was prepared from a single sheep spleen by standard procedures. The DNA was partially cleaved with *Sau*3A and size-fractionated on gradients of 10% to 40% (w/v) sucrose. Fractions containing DNA fragments of 14 to 20 kb were pooled. EMBL3 (Frischauf *et al.*, 1983) was digested with *Bam*HI and *Eco*RI, and precipitated with spermine (Hoopes & McClure, 1981). This procedure precipitates fragments greater than 100 bp in length and leaves the short *Bam*HI–*Eco*RI segment in solution, obviating the need to physically purify phage arms. The preparation was ligated with size-fractionated, *Sau*3A-cleaved sheep DNA, packaged and plated.

(b) *Isolation of BLG clones*

The unamplified library was plated on 22 cm × 22 cm plates at a plaque density of about 60,000/plate. Approximately 500,000 clones were screened. Plaque lifts were performed as described by Benton & Davis (1977). The library was probed by hybridization to the BLG plasmid p931 (Gaye *et al.*, 1986) essentially as described by Maniatis *et al.* (1978). Six positively hybridizing plaques were isolated and DNA was purified from 4 of them.

(c) *Cloned DNA*

The phage clones were mapped by cleavage with a variety of restriction endonucleases using standard gel electrophoresis and Southern blotting techniques. Subcloning into plasmid and M13 phage vectors was performed according to standard recombinant DNA procedures (Maniatis *et al.*, 1982). DNA sequencing was performed using the dideoxy method, essentially as described by Sanger *et al.* (1977) and Anderson *et al.* (1980).

3. Results and Discussion

(a) *Isolation and characterization of ovine BLG genes*

A library of partially *Sau*3A-digested sheep spleen DNA was constructed in the bacteriophage lambda vector EMBL3 (Frischauf *et al.*, 1983). Approximately 500,000 recombinants from the unamplified library were screened using the BLG cDNA plasmid p931 (Gaye *et al.*, 1986) as hybridization probe. Six clones were identified and isolated by plaque purification. Restriction maps were constructed for four of the clones, using six restriction enzymes in a combination of single and double digests (Fig. 1). After Southern blotting, filters were hybridized to p931 and specific 5′ and 3′ fragments derived from this plasmid. These experi-

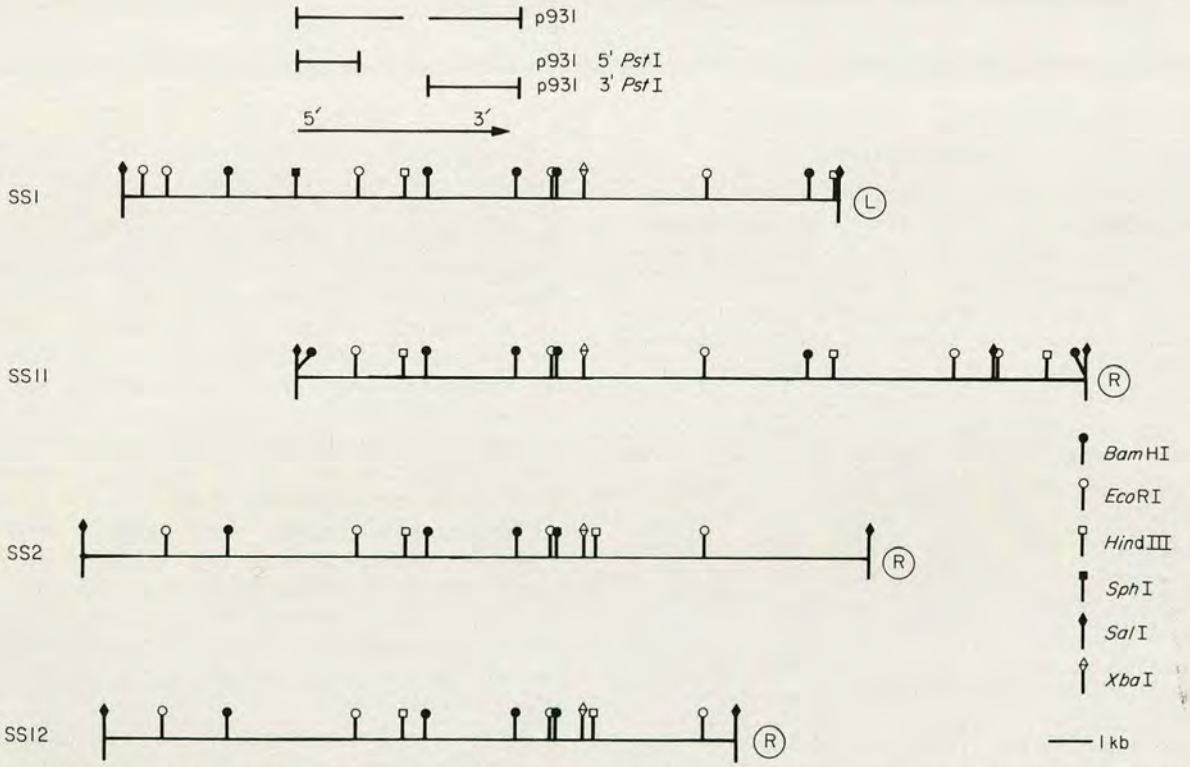


Figure 1. Restriction maps of BLG clones. The clones are aligned using common restriction enzyme sites. |—|, indicates the extent of hybridization of p931 and specific 5′ and 3′ *Pst*I fragments isolated from this clone. L and R refer to left and right phage arms.

ments defined the limits of hybridization and the orientation of BLG cDNA sequences within the genomic clones.

The clones contained inserts varying from 14 kb to 17·5 kb. Clones SS1 and SS11 have identical restriction maps over an overlapping region of 10·4 kb. Similarly, the insert of SS12 is identical with a segment of at least 11·7 kb within SS2. Although the SS1/SS11 clones are very similar to the SS2/SS12 clones, and can be aligned over a distance of at least 7·6 kb, there are some notable differences. Within the region in which the four clones overlap, SS2/SS12 are distinguished from SS1/SS11 by the presence of a *Hin*dIII site 1·9 kb from the 3′ end of the BLG gene. In other regions, in which only two or three of the clones overlap, an additional four restriction enzyme site differences are observed. One of these differences (the presence of an *Sph*I site in SS1) is in a region where three of the four clones can be aligned by virtue of common restriction sites on both sides of the *Sph*I site. It may, therefore, represent a point mutation, although a very small insertion/deletion cannot be ruled out. The three other restriction site differences noted are not flanked by common sites, so it is not possible to say whether these differences are most likely due to point mutations or insertion/deletion events.

The independent isolation of four very similar clones is strong evidence that no major cloning artefacts have occurred. SS1 and SS11 appear to contain identical overlapping chromosomal regions, as do SS2 and SS12. These data indicate that the two types of clone represent two similar, but non-identical, BLG genes. Southern blotting of sheep genomic DNA, followed by hybridization to BLG cDNA probes, detects restriction fragments predicted from the maps presented in Figure 1 at an intensity consistent with a copy number of 1 (Gaye *et al.*, 1986; and our unpublished results). This suggests that SS1/SS11 and SS2/SS12 may represent two BLG alleles. Polymorphic variation is found in at least two positions of ovine BLG, at amino acid positions +20 (His/Tyr) and +85 (Asp/Asn) (Kolde & Braunitzer, 1983*b*; Gaye *et al.*, 1986).

(b) *Gene structure of SS1*

Two subclones of SS1 were constructed by inserting the 4 kb *Bam*HI–*Hin*dIII fragment and the 4 kb *Hin*dIII–*Xba*I fragment, corresponding to 5′ and 3′ segments of the BLG gene (see Fig. 1) into the plasmid vector pPoly1 (Lathe *et al.*, 1987). These two subclones, named pSS1BH and pSS1HX, were mapped using a variety of restriction enzymes, a number of which were chosen because the presence of their cleavage sites in BLG exons was predicted from the published sequence of ovine BLG mRNA (Gaye *et al.*, 1986). Fragments generated by these enzymes were cloned into compatible restriction sites in the M13 vectors tg130 and tg131 (Kieny *et al.*, 1983). DNA sequencing of these clones identified fragments containing BLG exons. Furthermore, alignment of these sequences with the published BLG mRNA sequence allowed the precise determination of exon/intron boundaries and

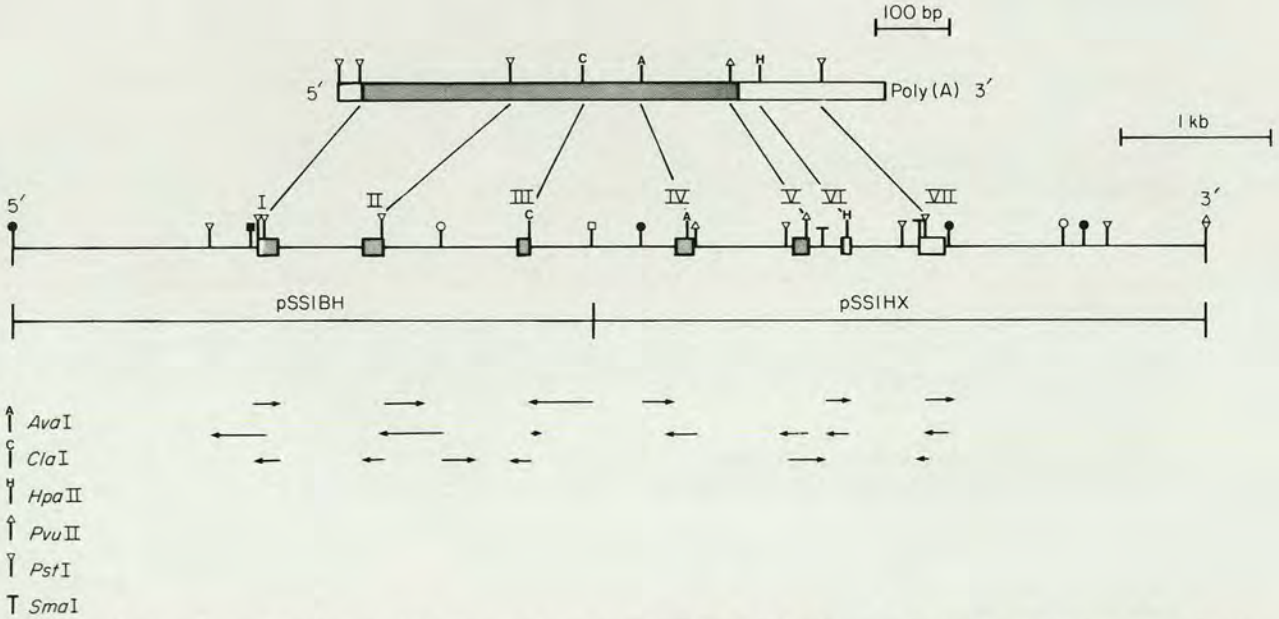


Figure 2. Structure of BLG gene SS1. The relationship of BLG mRNA (Gaye *et al.*, 1986) to SS1 is indicated using common restriction enzyme sites. The exons are shown as open boxes; coding regions are shaded. Symbols for restriction sites are the same as in Fig. 1; additional restriction sites used in the construction of M13 clones and for positioning exons are indicated; complete maps for these sites are not presented. Two subclones of SS1, pSS1BH and pSS1HX, are shown. The extent and direction of DNA segments sequenced is indicated by arrows.


Exon I (136 bp)
agccaccccgggcctaggatgagccaagtgggattccgggaaccgcntggctgggggccagcccgggctggctggcctgc
*
atgcgcctcctgtataaggccccaagcctgcctgtctcagccctccACTCCCTGCAGAGCTCAGAAGCACGACCCCAGCT
T
-18 -1 +1
MetLysCysLeuLeuLeuAlaLeuGlyLeuAlaLeuAlaCysGlyValGlnAlaIleIleValThrGlnThrMe
GCAGCCATGAAGTGCCTCCTGCTTGCCCTGGGCCTGGCCCTCGCCTGTGGCGTCCAGGCCATCATCGTCACCCAGACCAT
10
tLysGlyLeuAspIleGlnLys
GAAAGGCCTGGACATCCAGAAGgttcgagggt
Exon II (140 bp)
20 30
ValAlaGlyThrTrpHisSerLeuAlaMetAlaAlaSerAspIleSerLeuLeuAspAlaGlnSerAlaP
ccctctccagGTGGCGGGGACTTGGCACTCCTTGGCTATGGCGGCCAGCGACATCTCCCTGCTGGATGCCCAGAGTGCCC
Tyr
40 50 60
roLeuArgValTyrValGluGluLeuLysProThrProGluGlyAsnLeuGluIleLeuLeuGlnLysTr
CCCTGAGAGTGTACGTGGAGGAGCTGAAGCCCACCCCCGAGGGCAACCTGGAGATCCTGCTGCAGAAATGgtgggcgtct
Exon III (74 bp)
70 80
pGluAsnGlyGluCysAlaGlnLysLysIleIleAlaGluLysThrLysIleProAlaValPheLysIle
tgtctttcagGGAGAACGGCGAGTGTGCTCAGAAGAAGATTATTGCAGAAAAAACCAAGATCCCTGCGGTGTTCAAGATC
AspA
GATGgtgagtccgg
Asn
Exon IV (111 bp)
90 100
laLeuAsnGluAsnLysValLeuValLeuAspThrAspTyrLysLysTyrLeuLeuPheCysMetGluAs
ccgcgtccagCCTTGAATGAGAACAAAGTCCTTGTGCTGGACACCGACTACAAAAAGTACCTGCTCTTCTGCATGGAAAA
110 120
nSerAlaGluProGluGlnSerLeuAlaCysGlnCysLeuV
CAGTGCTGAGCCCGAGCAAAGCCTGGCCTGCCAGTGCCTGGgtgggtgcca
Exon V (105 bp)
130 140
alArgThrProGluValAspAsnGluAlaLeuGluLysPheAspLysAlaLeuLysAlaLeuProMetHi
tgccccatagTCAGGACCCCGGAGGTGGACAACGAGGCCCTGGAGAAATTCGACAAAGCCCTCAAGGCCCTGCCCATGCA
150
sIleArgLeuAlaPheAsnProThrGlnLeuGluG
CATCCGGCTTGCCTTCAACCCGACCCAGCTGGAGGgtgacgaccc
Exon VI (42 bp)
160
lyGlnCysHisValEnd
tcccccacagGGCAGTGCCACGTCTAGGTGAGCCCCTGCCGGTGCCTCTGGGgtaagctgct

Fig. 3.

Exon VII (180 bp)

ccattttcagGGCCCGGGAGCCTTGGACTCCTCTGGGGACAGACGACGTCACCACCGCCCCCCCCCCATCAGGGGGACTA

GAAGGGACCAGGACTGCAGTCACCCTTCCTGGGACCCAGGCCCCTCCAGGCCCCTCCTGGGGCTCCTGCTCTGGGCAGCT

TCTCCTTCACCAATAAAGGCATAAACCTGTgctctcccttctgagtctttgctggacgacgggcaggggt

Figure 3. DNA sequence of BLG gene SS1. Exon sequences are shown in upper case and flanking sequences in lower case. The predicted protein sequence is shown immediately above the DNA sequence. CAT, TATAA and AATAAA signals are underlined. The putative mRNA cap site is shown (*). Differences between SS1 and the published sequence of BLG mRNA (Gaye *et al.*, 1986) are indicated on the line below the sequence; (.) indicates a gap in the alignment.

therefore the genomic organization of the ovine BLG gene (Fig. 2). Exons were positioned within the BLG transcription unit by reference to the position of the restriction sites in introns and to sites held in common between the BLG cDNA sequence and SS1, and by probing appropriate restriction digests with exon-specific M13 clones and defined 5′ and 3′ fragments prepared from p931 (data not shown). The length of each exon was precisely determined by DNA sequencing and the length of the introns estimated from restriction fragment lengths. The BLG gene spans 4·9 kb and comprises seven exons (Fig. 2). Exons I to III are contained within the plasmid subclone pSS1BH and exons IV to VII within pSS1HX. Introns range in size from 0·2 kb (intron 5) to 1·0 kb (intron 3) and together they account for approximately 84% of the gene.

(c) *Sequence analysis of SS1*

The nucleotide sequence of exons I to VII and the translated amino acid sequence are shown in Figure 3. The gene contains an open reading frame of 540 nucleotides coding for a 180 amino acid residue pre-protein comprising the 18-residue signal peptide (Mercier *et al.*, 1978) followed by the 162 residues of the mature protein. Translation is initiated in exon I and terminated in exon VI. Exon VII is entirely non-coding. The translated sequences of SS1 are identical to those published for BLG mRNA (Gaye *et al.*, 1986). The histidine codon (CAC) at position 20 indicates that SS1 encodes the B, rather than the A (Tyr) variant of BLG (Gaye *et al.*, 1986; Kolde & Braunitzer, 1983*b*).

We have positioned an mRNA cap site at the A residue (.ccctcccACTCCCT..) 40 bp upstream from the A of the initiating methionine codon. Primer extension studies on BLG mRNA have unambiguously defined the 5′-terminal sequences up to the C residue that immediately follows this A (Gaye *et al.*, 1986). In this respect, the presumed cap site of SS1 is similar to that of many other eukaryote mRNAs, which have been shown to comprise an A surrounded by pyrimidines (Breathnach & Chambon, 1981). A TATA box, believed to control the accuracy of transcription through site-specific initiation, and a possible CAT consensus signal (Benoist *et al.*, 1980) are located 34 bp and 102 bp upstream, respectively, from this A (Table 1). The untranslated leader sequence of SS1 differs at one position when compared to the published BLG mRNA sequence (a C in place of T 14 bp downstream from the designated cap site).

Table 1 tabulates the donor and acceptor sites of the six introns of SS1. All 12 sites accord with the GT/AG rule and show good agreement with the consensus sequences derived by Breathnach & Chambon (1981). In addition, we have tabulated a sequence conforming to the consensus sequence CTGAC (Keller & Noon, 1984) present in each intron between −20 and −55 bp from the acceptor site boundary. In this consensus, the A is always present and is known in some cases to participate in the formation of the branch-point of the lariat splicing intermediate (Ruskin *et al.*, 1984). Five of the six introns of SS1 contain a sequence that conforms in at least four out of five positions.

The 3′ untranslated sequences of SS1 in exon VI and in the non-coding exon VII are identical to those published for BLG mRNA, except that SS1 contains an extra C in a run of C residues that starts 49 bp from the beginning of exon VII. The polyadenylation signal (AATAAA) in exon VII is found 13 bp upstream from the site of poly(A) addition.

(d) *SS1 is a functional BLG gene*

All the available evidence indicates that SS1 is a true gene, i.e. one that is transcribed and translated, rather than a pseudogene. It contains a correctly positioned TATA box, mRNA cap site and 3′ polyadenylation signal. All of the designated donor and acceptor splice sites conform to the established consensus sequences, particularly the AG/GT rule (Breathnach & Chambon, 1981). Finally, the gene contains a 180-codon open reading frame, identical with the published sequence of ovine BLG mRNA (Gaye *et al.*, 1986), which encodes a protein whose sequence agrees with the

Table 1
DNA sequence signals present in SS1

Signal	Gene	Sequence	Position
Transcription initiation	BLG	A G C C A A G T G	−102
	Consensus[a]	G G Y C A A T C T	−80
	BLG	C T G T A T A A G G C C	−34
	Consensus[a]	G N G T A T A W A W N G	−30
	BLG	C A C T C C	+1
	Consensus[b]	C A N Y Y Y	+1
Donor/lariat/acceptor splice sites			
Intron 1		G T T C G A / C C G A C / C C C T C T C C A G	
Intron 2		G T G G G C / C T G A T / T G T C T T T C A G	
Intron 3		G T G A G T / C T C A C / C C G G G T C C A G	
Intron 4		G T G G G T / C T G A C / T G C C C C A T A G	
Intron 5		G T G A C G / C T G A C / T C C C C C A C A G	
Intron 6		G T A A G C / C A C A G / C C A T T T T C A G	
Consensus[a,c]		G T R A G T / C T R A Y / Y Y Y Y Y Y N C A G	
Poly(A) addition signals	BLG	A A T A A A	
	Consensus[a]	A A T A A A	
Translation initiation	BLG	C A G C C A T G	+41
	Consensus[d]	C C R C C A T G	
Translation termination	BLG	T A G	

Consensus signals were taken from [a] Breathnach & Chambon (1981), [b] Bucher & Trifonov (1986). [c] Keller & Noon (1984), and [d] Kozak (1984).

known amino acid sequence of ovine BLG (Kolde & Braunitzer, 1983*a*).

Strong evidence for the functional integrity of SS1 has come from experiments with transgenic mice. Mice carrying either the entire 15·5 kb *Sal*I insert or the 10·8 kb *Xba*I–*Sal*I fragment of SS1 (see Fig. 1) specifically express the gene in the lactating mammary gland (Simons *et al.*, 1987). The mouse transcripts are the same size as sheep BLG mRNA and are efficiently translated: a number of lines of transgenic mice have been produced that contain high concentrations of ovine BLG in their milk. The BLG produced in these mice appears, when examined by SDS/polyacrylamide gel electrophoresis and Western blots, identical with sheep BLG.

(e) *The relationship between BLG and RBP*

The gene organization of rat RBP has been published and compared to the three-dimensional structure of human RBP (Laurent *et al.*, 1985). This comparison shows that the sequences encoded by individual exons of the RBP gene closely correspond to discrete tertiary structural elements. Since BLG and RBP have a similar three-dimensional structure, we have compared the organization of their respective genes with particular reference to protein structure.

Figure 4 shows the exons of each gene schematically aligned with the tertiary folds of the corresponding protein. In this comparison, the gene for rat RBP has been aligned with human RBP and

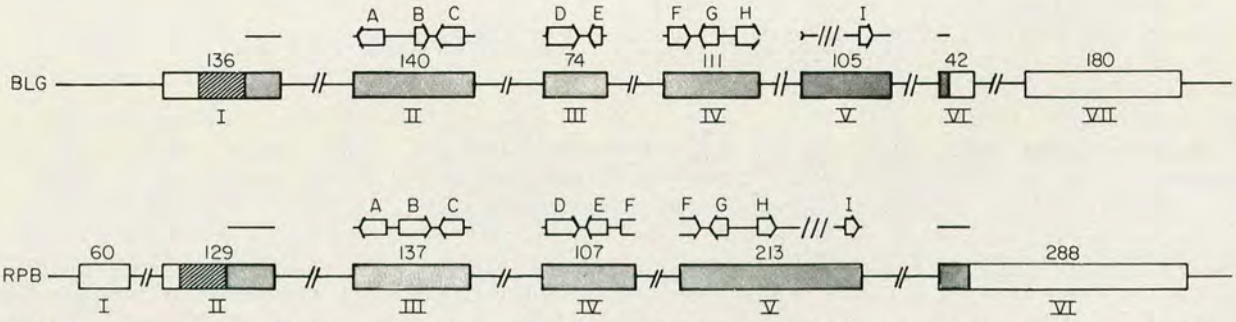


Figure 4. Comparison of 3-dimensional elements of BLG and RBP with their corresponding gene structures. Exons are shown as an open box and protein coding regions are shaded (the signal peptide region is hatched). Exon sizes (in bp) are shown directly above. Corresponding 3-dimensional elements of each protein are indicated above each gene (after Laurent *et al.*, 1985). β-Sheets are lettered A to I and their antiparallel nature is indicated by arrows. N-terminal and C-terminal coils and connecting loops are shown as horizontal lines and the α-helix is represented by 3 slashes.

the gene for ovine BLG (SS1) aligned with bovine BLG. Rat and human RBP are closely related (Sundelin *et al.*, 1984) and will have similar tertiary folding. Similarly, we have assumed that the structure of bovine BLG (Papiz *et al.*, 1986) is a good prediction of ovine BLG, since the two proteins differ at only 5 out of 162 positions (Braunitzer *et al.*, 1972: Kolde & Braunitzer, 1983*b*).

Both BLG and RBP are characterized by eight antiparallel β-strands (A to H), which comprise the β-barrel core, followed by a three-turn α-helix and a further stretch of β-strand (I) close to the C terminus. In BLG, strand I is involved in the formation of the dimer by interacting with the dyad-related strand in the other subunit.

The exons encoding these elements show distinct homologies when the two gene structures are compared in this manner. Exon I of BLG and exon II of RBP encode the signal peptide and N-terminal coil. Exon II of BLG and exon III of RBP are very similar in size 140 *versus* 137 nucleotides) and both encode the first three β-strands (A, B and C) of the β-barrel. Both exons terminate at a position encoding the middle of a reverse turn of a β-sheet. Exon III of BLG is 74 bp in length and corresponds almost exactly to β-strands D and E; it also terminates at a position corresponding to the middle of a reverse turn. Exon IV of RBP also encodes strands D and E but is 33 bp longer and encodes the first part of an extended strand F. In this case, the exon does not terminate at sequences encoding a reverse turn but at a position that is just before a β-bend.

Exon IV of BLG encodes the sequences that comprise strands F, G and H of the β-barrel, and exon V encodes the three-turn α-helix and strand I. In contrast, these five three-dimensional elements are encoded entirely by exon V of RBP. The correspondance of the three-dimensional elements encoded, and the fact that the combined size of BLG exons IV and V (216 bp) is just 3 bp longer than RBP exon V (213 bp), suggests that this difference may have arisen as a result of the insertion or deletion of an intron.

The translation termination codon for both proteins is in exon VI; the short C-terminal amino segment is disordered, and in both cases contains a cysteine residue that forms a disulphide bond with a cysteine residue present in β-strand D.

Both BLG and RBP seem to agree with the general proposal of Craik *et al.* (1982), that exon junctions map to the protein surface. Furthermore, the similarity of their tertiary folding is reflected by their gene organization, since both genes encode homologous three-dimensional elements within corresponding exons.

(f) *Comparison with genes encoding other secretory proteins*

The rodent urinary proteins, rat alpha-2-urinary globulin (α2-UG) and mouse major urinary protein (MUP), human apolipoprotein D (apo-D), and rat and human alpha-1-acid glycoprotein (α1-AGP) have limited amino acid sequence homology with BLG and RBP (Unterman *et al.*, 1981; Drayna *et al.*, 1986; J. O. Bishop, personal communication). The structures of the genes encoding these proteins

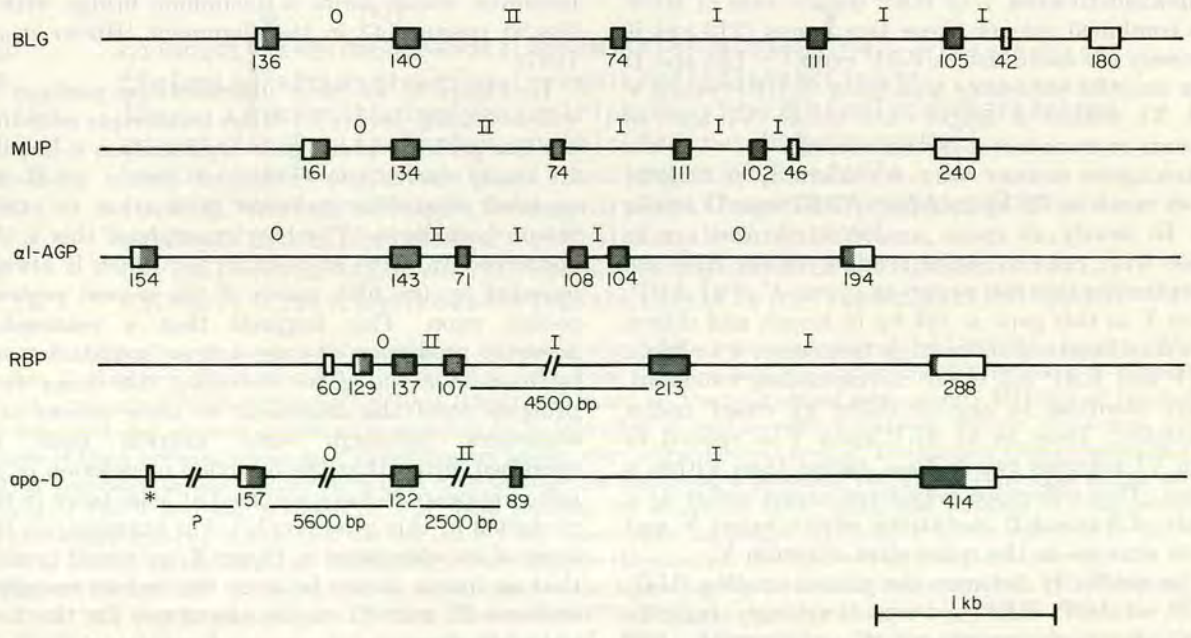


Figure 5. Comparison of gene structures. Exons are shown as open boxes, and their coding regions are shaded. The 5 genes are aligned at their second protein-coding exon. The phase of each intron is indicated: 0, splicing between 2 codons; I, splicing between 1/3 (5′) and 2/3 (3′) of a codon and II, *vice versa* to I. BLG, ovine beta-lactoglobulin; MUP, murine major urinary protein (Clark *et al.*, 1984); α1-AGP, human alpha-1-acid glycoprotein (Dente *et al.*, 1987); RBP, rat retinol binding protein (Laurent *et al.*, 1985); apo-D, human apolipoprotein D (Drayna *et al.*, 1987).

intron–exon junction. An association of splice junction position with variation in protein structure has been noted by Craik *et al.* (1983) for the serine proteases and dihydrofolate reductases.

(h) *A common three-dimensional structure?*

The calculated homology between BLG and RBP based on the alignment in Figure 6 is 15·1%, which is about average for all pair-wise comparisons of the nine proteins. A comparison (L. Sawyer, personal communication) of tertiary structures of BLG and RBP using the method of Rao & Rossmann (1973) and Rossmann & Argos (1975), which generates an alignment of amino acids based upon their proximity in space, gave an alignment almost identical to that presented in Figure 6. As noted in the previous section, some amino acid positions are strongly conserved throughout the family and may be functionally significant. In particular, the conserved Trp at position 27 has been implicated in binding retinol in BLG (Papiz *et al.*, 1986). The similarity of the folding of BLG and RPB shows that highly related amino acid sequences are not a prequisite for a similar three-dimensional structure. The three-dimensional structure of INCYN has been published and shown to have a β-barrel configuration similar to that of BLG and RBP (Huber *et al.*, 1987; Holden *et al.*, 1987). In addition, a comparison of BLG, MUP and α2-UG using the modified program suite PREDICT (Eliopoulos *et al.*, 1982; Sawer *et al.*, 1986) gave similar predictions for the occurrence of turns or coils, extended β-sheet and α-helix structure (L. Sawyer, personal communication).

(i) *Functional properties*

We have described a diverse group of proteins that are probably evolutionarily related. If so, this gene family is ancient, as examples are found in vertebrates and invertebrates.

In Table 2, we have summarized some of the properties and functions of these proteins (where known). They are relatively small (160 to 189 amino acid residues) and are encoded by single genes, e.g. BLG (Aschaffenburg & Drewry, 1957; this paper) or multigene families, e.g. MUP (Clark *et al.*, 1982). The genes encoding them are expressed in a variety of secretory tissues and the corresponding proteins are secreted (or presumed to be) into a wide variety of bodily fluids including blood serum, milk, saliva,

Table 2

Properties of related secretory proteins

Protein	Size	Tissue	Fluid	Properties
BLG	162	Mammary gland	Milk	Binds retinol, and a gut receptor; possible involvement in vitamin A transport to young (1, 2)
MUP (α2-UG)	162	Liver, salivary lachrymal, preputial glands	Serum, urine, saliva*, tears*, seminal fluid*	Possible binding of pheromones (3, 4, 5)
α1-AGP	187	Liver	Serum	Mediates inflammatory response; can bind progesterone (6, 7, 8, 9, 10)
RBP	183	Liver	Serum	Binds and transports retinol in the serum, in association with transthyretin (11, 12)
apo-D	169	Adrenal, kidney, pancreas, liver, intestine	Serum, gut secretions*	Binds lecithin: cholesterol acyltransferase; possibly involved in cholesterol transport (13)
HCHU (α1-MG)	183	Liver	Serum, urine cerebrospinal fluid	Binds yellow-brown retinoid, and also binds IgA; involved in mediating neutrophil chemotaxis (14, 15, 16)
BG	160	Olfactory epithelium	Nasal mucus*	Binds odorants; probably involved in presentation of odorants to receptors on neurones (17)
ESP	165	Epididymis	Epididymal luminal fluid*	Binds sperm membrane (18)
INCYN	189	Larval fat body?	Hemolymph	Binds chromophores called biliverdins; involved in colour camouflage (19)
PP14 (α2-PEG)	?	Placenta, secretory endometrium	Amniotic fluid	Synthesized during mid to late-luteal and early pregnancy stages (20, 21)

Column 2 shows the number of amino acid residues making up the mature polypeptide. Also listed are the tissues in which the genes are expressed, and the fluids in which the proteins have been found. The fluids marked with an asterisk are presumed to contain the protein because the presence of mRNA has been noted in the appropriate tissue. The final column lists some properties and functions.

MUP and α2-UG are listed together as they are equivalent rodent urinary proteins (see the text). Alpha-1-microglobulin (α1-MG) has been listed under HCHU, since they are probably identical (Akerstrom *et al.*, 1979).

The human pregnancy-associated endometrial α2-globulin (α2-PEG), which has been shown to be immunochemically similar to placental protein 14 (PP14; Bell & Bohn, 1986) has been partially sequenced. A comparison of the amino-terminal 38 residues shows a 45% homology with BLG (Huhtala *et al.*, 1987; Bell *et al.*, 1987). α2-PEG would appear to be a member of this family of proteins.

The data were compiled from the following sources: (1) Futterman & Heller (1972); (2) Papiz *et al.* (1986); (3) Shaw *et al.* (1983); (4) Held & Gallagher (1985); (5) Clark *et al.* (1985); (6) Liao *et al.* (1985); (7) Baumann & Held (1981); (8) Baumann *et al.* (1983); (9) Wagh *et al.* (1969); (10) Ganguly *et al.* (1967); (11) Peterson *et al.* (1973); (12) Laurent *et al.* (1985); (13) Drayna *et al.* (1986); (14) Tehler *et al.* (1978); (15) Tejler & Grubb (1976); (16) Mendez *et al.* (1986); (17) Lee *et al.* (1987); (18) Brooks *et al.* (1986); (19) Riley *et al.* (1984); (20) Huhtala *et al.* (1987); (21) Bell *et al.* (1987). 'Note added in proof'. Purpurin, a protein found in the neural retina which binds retinol has been shown recently to be another member of this group of proteins (Berman, P., Gray, P., Chen, E., Keyser, K., Ehrlich, D., Karten, H., LaCorbiere, M., Esch, F. & Schubert, D. (1987). *Cell*, **51**, 135–142).

seminal fluid, amniotic fluid, tears and nasal mucus.

RBP binds retinol and transports it in the serum. BLG also binds retinol, and it has been proposed that it may function in its transport (Papiz *et al.*, 1986). HCHU has been implicated in the binding of retinoid complexes (Pervaiz & Brew, 1985), apo-D is thought to bind lecithin and cholesterol or cholesteryl esters (Drayna *et al.*, 1986) and INCYN binds biliverdins (Cherbas, 1973).

The binding of small hydrophobic molecules may be a common feature of many of these proteins. The possibility that frog BG solubilizes and concentrates odorants in mucus (Lee *et al.*, 1987) is consistent with this proposal. Similarly, a protein called odorant binding protein has been isolated from bovine nasal mucosa (Pevsner *et al.*, 1986) and shown to bind a variety of small hydrophobic odorant molecules. Its molecular weight of 19,000 is consistent with it being a member of this family of proteins. MUP may also have a role in the binding and presentation of odorants, as male mouse urine has been shown to contain protein agents that have a dramatic pheromonal effect when administered to young females (Vandenburgh *et al.*, 1975).

In conclusion, we have reported the gene structure of BLG and shown that it is similar to the genes encoding MUP, α1-AGP, RBP and apo-D. The proteins appear to be members of an ancient gene family comprising a wide variety of secretory proteins characterized by a few highly conserved amino acid positions and, very often, by the ability to bind small hydrophobic molecules. Three of these proteins are now known to have similar three-dimensional structures.

We thank Dr J.-C. Mercier for his gift of plasmid p931. we are grateful to A. L. Archibald, J. O. Bishop, S. Harris, L. Sawyer, P. Simons, J.-L. Vilotte and C. B. A. Whitelaw for comments on the manuscript and for helpful discussion.

References

Akerstrom, B., Nilsson, K., Berggard, B. & Berggard, I. (1979). *J. Immunol.* **122**, 2516–2520.

Anderson, S., Gait, M. J., Mayol, L. & Young, I. (1980). *Nucl. Acids Res.* **8**, 1731–1743.

Aschaffenburg, R. & Drewry, J. (1957). *Biochem. J.* **65**, 273–277.

Baumann, H. & Held, W. A. (1981). *J. Biol. Chem.* **256**, 10145–10155.

Baumann, H., Firestone, G. L., Burgess, T. L., Gross, K. W., Yamamoto, K. R. & Held, W. A. (1983). *J. Biol. Chem.* **258**, 563–570.

Bell, S. C. & Bohn, H. (1986). *Placenta*, **7**, 283–294.

Bell, S. C., Keyte, J. W. & Waites, G. T. (1987). *J. Clin Endocrinol. Metabol.* in the press.

Benoist, C., O'Hare, K., Breathnach, R. & Chambon, P. (1980). *Nucl. Acids Res.* **8**, 127–142.

Benton, W. D. & Davis, R. W. (1977). *Science*, **196**, 180–182.

Braunitzer, G., Chen, R., Schrank, B. & Stangl, A. (1972). *Hoppe-Seyler's Z. Physiol. Chem.* **353**, 832–834.

Breathnach, R. & Chambon, P. (1981). *Annu. Rev. Biochem.* **50**, 349–383.

Bucher, P. & Trifonov, E. N. (1986). *Nucl. Acids Res.* **14**, 10009–10026.

Brooks, D. E., Means, A. R., Wright, E. J., Singh, S. P. & Tiver, K. K. (1986). *J. Biol. Chem.* **261**, 4956–4961.

Brooks, D. E. (1987). *Biochem. Int.* **14**, 235–240.

Cherbas, P. K. (1973). Ph.D. thesis, Harvard University.

Clark, A. J., Clissold, P. M. & Bishop, J. O. (1982). *Gene*, **18**, 221–230.

Clark, A. J., Clissold, P. M., Al Shawi, R., Beatie, P. & Bishop, J. (1984). *EMBO J.* **3**, 1045–1052.

Clark, A. J., Ghazal, P., Bingham, R. W., Barrett, D. & Bishop, J. O. (1985). *EMBO J.* **4**, 3159–3165.

Cogan, U., Kopelman, M., Mokady, S. & Shinitzky, M. (1976). *Eur. J. Biochem.* **6**, 71–78.

Craik, C. S., Sprang, S., Fletterick, R. & Rutter, W. J. (1982). *Nature (London)*, **299**, 180–182.

Craik, C. S., William, J. R. & Fletterick, R. (1983). *Science*, **220**, 1125–1129.

Dayhoff, M. O. (1969). Editor of *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Silver Springs, MD.

Dente, L., Pizza, M. G., Metspalu, A. & Cortese, R. (1987). *EMBO J.* **6**, 2289–2296.

Devereux, J. & Haeberli, P. (1984). Program Library of the University of Wisconsin Genetics Computer Group.

Drayna, D., Fielding, C., Mclean, J., Baer, B., Castro, G., Chen, E., Comstock, L., Henzel, W., Kohr, W., Rhee, L., Wion, K. & Lawn, R. (1986). *J. Biol. Chem.* **261**, 16535–16539.

Drayna, D. T., McLean, J. W., Wion, K. L., Trent, J. M., Drabkin, H. A. & Lawn, R. M. (1987). *DNA*, **6**, 199–204.

Eliopoulos, E. E., Geddes, A. J., Brett, M., Papin, D. J. C. & Findlay, J. B. C. (1982). *Int. J. Biol. Macromol.* **4**, 263–268.

Frischauf, A. M., Lehrach, H., Poutska, A. & Murray, N. (1983). *J. Mol. Biol.* **170**, 827–842.

Fugate, R. D. & Song, P.-S. (1980). *Biochim. Biophys. Acta*, **625**, 28–42.

Futterman, S. & Heller, J. (1972). *J. Biol. Chem.* **247**, 5168–5172.

Ganguly, M., Carnighan, R. H. & Westphal, U. (1967). *Biochemistry*, **6**, 2803–2814.

Gaye, P., Hue-Delahaie, D., Mercier, J.-C., Soulier, S., Vilotte, J.-L. & Furet, J.-P. (1986). *Biochimie*, **68**, 1097–1107.

Godovac-Zimmerman, J., Conti, A., Liberatori, J. & Braunitzer, G. (1985). *Hoppe-Seyler's Z. Physiol. Chem.* **366**, 431–434.

Held, W. A. & Gallagher, J. F. (1985). *Biochem. Genet.* **23**, 281–290.

Holden, H. M., Rypniewski, W. R., Law, J. H. & Rayment I. (1987). *EMBO J.* **6**, 1565–1570.

Hoopes, B. C. & McClure, R. R. (1981). *Nucl. Acids Res.* **9**, 5493–5504.

Huber, R., Schneider, M., Epp, O., Mayr, I., Messerschmidt, A. & Pflugrath, J. J. (1987). *J. Mol. Biol.* **195**, 423–434.

Huhtala, M.-L., Seppala, M., Narvanen, A., Palomaki, P., Julkunen, M. & Bohn, H. (1987). *Endocrinology*, **120**, 2620–2622.

Jenness, R. (1982). In *Developments in Dairy Chemistry* (Fox, P. F., ed.), vol. 1, pp. 83–114, Applied Sciences Publishers, London and New York.

Kaumeyer, F., Polazzi, J. O. & Kotick, M. P. (1986). *Nucl. Acids Res.* **14**, 7839–7850.

Keller, E. B. & Noon, W. A. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 7417–7420.

Kieny, M. P., Lathe, R. & Lecocq, J. P. (1983). *Gene*, **26**, 91–99.

Kolde, H. J. & Braunitzer, G. (1983*a*). *Milschwissenschaft*, **38**, 18–20.
Kolde, H. J. & Braunitzer, G. (1983*b*). *Milschwissenschaft*, **38**, 70–72.
Kozak, M. (1984). *Nucl. Acids Res.* **12**, 857–872.
Lathe, R., Vilotte, J. L. & Clark, A. J. (1987). *Gene*, **57**, 193–201.
Laurent, B. C., Nilsson, M. H. L., Bavik, B. O., Jones, T. A., Sundelin, J. & Peterson, P. A. (1985). *J. Biol. Chem.* **260**, 11476–11480.
Lee, K. H., Wells, R. G. & Reed, R. R. (1987). *Science*, **235**, 1053–1056.
Liao, Y. J., Taylor, J. M., Vannice, J. L., Clawson, G. A. & Smuckler, E. A. (1985). *Mol. Cell. Biol.* **5**, 3634–3639.
Lopez. C., Grubb, A. & Mendez, E. (1981). *Biochem. Biophys. Res. Commun.* **103**, 919–925.
Lovrien, R. & Anderson, W. F. (1969). *Arch. Biochem. Biophys.* **131**, 139–144.
Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G.-K. & Efstratiadis, A. (1978). *Cell*, **15**, 687–701.
Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982). Editors of *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Mendez, E., Fernandez-Luna, J. L., Grubb. A. & Leyva-Cobian, F. (1986). *Proc. Nat. Acad. Sci., U.S.A.* **83**, 1472–1475.
Mercier, J.-C., Haze, G., Gaye, P. & Hue-Delahaie, D. (1978). *Biochem. Biophys. Res. Commun.* **68B**, 103–111.
Mercier, J.-C., Gaye, P., Soulier, S., Hue-Delahaie, D. & Vilotte, J. L. (1985). *Biochimie*, **67**, 959–971.
Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453.
Newcomer, M. E., Jones, T. A., Aqvist, J., Sundelin, J., Eriksson, U., Rask, L. & Peterson, P. A. (1984). *EMBO J.* **3**, 1451–1454.
Papiz, M. Z., Sawer, L., Eliopoulos, E. E., North, A. C. T., Findlay, J. B. C., Sivaprasadarao, R., Jones, T. A., Newcomer, M. E. & Kraulis, P. J. (1986). *Nature (London)*, **324**, 383–385.
Pervaiz, S. & Brew, K. (1985). *Science*, **228**, 335–337.
Peterson, P. A., Rask, L., Ostberg, L., Andersson, L., Kamwendo, F. & Pertoft, H. (1973). *J. Biol. Chem.* **248**, 4009–4022.
Pevsner, J., Sklar, P. B. & Snyder, S. H. (1986). *Proc. Nat. Acad. Sci., U.S.A.* **83**, 4942–4946.
Rao, S. T. & Rossmann, M. G. (1973). *J. Mol. Biol.* **76**, 241–256.
Riley, C. T., Barbeau, B. K., Keim, P. S., Kezdy, F. J., Heinrikson, R. L. & Law, J. H. (1984). *J. Biol. Chem.* **259**, 13159–13165.
Rossmann, M. G. & Argos, P. (1975). *J. Biol. Chem.* **250**, 7525–7532.
Ruskin, B., Krainer, A. R., Maniatis, T. & Green, M. B. (1984). *Cell*, **37**, 415–427.
Sanger, F., Nicklen, S. & Coulson, A. R. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 5463–5467.
Sawyer, L., Fothergill-Gilmore, L. A. & Russell, G. A. (1986). *Biochem. J.* **236**, 127–130.
Schmid, K., Burgi, W., Collins, J. H. & Nanno, S. (1974). *Biochemistry*, **13**, 2694–2697.
Shaw, P. D., Held, W. A. & Hastie, N. D. (1983). *Cell*, **32**, 755–761.
Simons, J. P., McClenaghan, M. & Clark, A. J. (1987). *Nature (London)*, **328**, 530–532.
Staden, R. (1982). *Nucl. Acids Res.* **10**, 2951–2961.
Sundelin, J., Laurent, B. C., Anundi, H., Tragardh, L., Larhammer, D., Peterson, P. A. & Rask, L. (1984). *J. Biol. Chem.* **259**, 6472–6484.
Tejler, L. & Grubb, A. O. (1976). *Biochim Biophys. Acta*, **439**, 82–94.
Tejler, L., Erikkson, S., Grubb, A. & Astedt, B. (1978). *Biochim. Biophys. Acta*, **542**, 506–514.
Unterman, R. D., Lynch, K. R., Nakhasi, H. L., Dolan, K. P., Hamilton, J. W., Cohn, D. V. & Feigelson, P. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 3478–3482.
Vandenburgh, J. G., Whitsett, J. M. & Lombardi, J. R. (1975). *J. Reprod. Fertil.* **43**, 515–523.
Wagh, P. V., Bornstein, I. & Winzler, R. J. (1969). *J. Biol. Chem.* **244**, 658–665.

Edited by P. Chambon

```
4166 ccggggagtttcaattatttcccaaaataagaactcaggtacaaagccatctttcaactatcacatcctgaaaacaaatggcagg
4251 tgacattttctgtgccgtagcagtcccactgggcattttcagggcccctgtgccaggggggcgcgggcatcggcgagtggaggct
4336 cctggctgtgtcagccggcccaggggggaggaagggacccggacagccagaggtgggggcaggctttccccctgtgacctgcaga
4421 cccactgcactgccctgggaggaaggggaggggaactaggccaaggggaagggcaggtgctctggagggcaagggcagacctgca
4506 gaccaccctggggagcagggactgacccccgtccctgccccatagTCAGGACCCCGGAGGTGGACAACGAGGCCCTGGAGAAATT
4591 CGACAAAGCCCTCAAGGCCCTGCCCATGCACATCCGGCTTGCCTTCAACCCGACCCAGCTGGAGGgtgagcacccaggccccgcc
4676 cttccccagggcaggagccacccggccccgggacgacctcctcccatggtgacccccagctccccaggcctcccaggaggaaggg
4761 gtggggtgcagcaccccgtgggggccccctccccacccctgccaggcctctcttcccgaggtgtccagtcccatcctgacccc
4846 ccatgactctccctcccccacagGGCAGTGCCACGTCTAGGTGAGCCCCTGCCGGTGCCTCTGGGgtaagctgcctgccctgccc
4931 cacgtcctgggcacacacatggggtaggggtcttggtggggcctgggaccccacatcaggccctggggtccccctgtgagaat
5016 ggctggaagctggggtcctcctggcgactgcagagctggctggccgcgtgccactcttgtgggtgacctgtgtcctggcctcac
5101 acactgacctcctccagctccttccagcagagctaaggctaagtgagccagaatggtacctaaggggaggctagcggtccttctc
5186 ccgaggaggggctgtcctggaaccaccagccatggagaggctggcaagggtctggcaggtgccccaggaatcacagggggccccc
5271 atgtccatttcagGGCCCGGGAGCCTTGGACTCCTCTGGGGACAGACGACGTCACCACCGCCCCCCCCCCATCAGGGGACTAGA
5356 AGGGACCAGGACTGCAGTCACCCTTCCTGGGACCCAGGCCCCTCCAGGCCCCTCCTGGGGCTCCTGCTCTGGGCAGCTTCTCCTT
5441 CACCAATAAAGGCATAAACCTGTgctctcccttctgagtctttgctggacgacgggcaggggtggagaagtggtgggaggggag
5526 tctggctcagaggatgacagcggggctgggatccagggcgtctgcatcacagtcttgtgacaactgggggcccacacacatcact
5611 gcggctctttgaaactttcaggaaccagggagggactcggcagagacatctgccagttcacttggagtgttcagtcaacacccaa
5696 actcgacaaaggacagaaagtggaaaatggctgtctcttagtctaataaatattgatatgaaactcaagttgctcatggatcaat
5781 atgcctttatgatccagccagccactactgtcgtatcaactcatgtacccaaacgcactgatctgtctggctaatgatgagagat
5866 tcccagtagagagctggcaagaggtcacagtgagaactgtctgcacacacagcagagtccaccagtcatcctaaggagatcagtc
5951 ctggtgttcattggaggactgatgttgaagctgaaactccaatgctttggccacctgatgtgaagagctgactcatttgaaaaga
6036 ccctgatgctgggaaagattgagggcaggaggagaagggagcgacagaggatgagatggttggatggcatcaccaacacaatgga
6121 catgggtttgggtggactccaggagttggtgatggacagggaggcctggcgtgctacggaagcggtttatggggtcacaaagact
6206 gagtgactgaactgagctgaactgaatggaaatgaggtatacagcaaagtggggatttttagataataagaatatacacataac
6291 atagtgtatactcatatttttatgcatacctgaatgctcagtcactcagtcgtatctgactctgtgacctatggaccgtagcctt
6376 ccaggtttcttctgtccacagaatttctccaaggcaagaatactggagtgggtagccatttcctcctccaggggatcctcccgacc
6461 cagggattgaaccggcatctcctgtattggcaggtggattctttaccactgtgccaccagggaagcccgtgttactctctatgtc
6546 ccacttaattaccaaagctgctccaagaaaaagcccctgtgccctctgagcttcccggcctgcagagggtggtgggggtagactg
6631 tgacctgggaacaccctcccgcttcaggactcccgggccacgtgacccacagtcctgcagacagccgggtagctctgctcttcaa
6716 ggctcattatctttaaaaaaaactgaggtctattttgtgacttcgctgccgtaacttctgaacatccagtgcgatggacaggacc
6801 tcctccccaggcctcagggcttcagggagccagccttcacctatgagtcaccagacactcgggggtggccccgccttcagggtg
6886 ctcacagtcttcccatcgtcctgatcaaagagcaagaccaatgacttcttaggagcaagcagacacccacaggacactgaggttc
6971 accagagctgagctgtccttttgaacctaaagacacacagctctcgaaggttttctctttaatctggatttaaggcctacttgcc
7056 cctcaagagggaagacagtcctgcatgtccccaggacagccactcggtggcatccgaggccacttagtattatctgaccgcaccc
7141 tggaattaatcggtccaaactggacaaaaaccttggtgggaagtttcatcccagaggcctcaaccatcctgctttgaccaccctg
7226 catcttttttttcttttatgtgtatgcatgtatatatatatatatttttttttttttcatttttggctgtgctggctgttcgt
7311 tgcagttcggtgcgcaggcttctctctagtttctctctagtcttctcttatcacagagcagtctctaga  7379
```

Fig 1. The 7379bp of genomic sequence encompassing the ovine BLG gene. Upper case sequence corresponds to the transcribed exons that comprise the mature message (3,6). Underlined at 5938-6231 and 6328-6444 are sequences with homology to previously described artiodactyl Alu-like repeat sequences (7,8 respectively).

References

1. Jenness, R. (1982) In Fox, P.F. (ed), Developments in Dairy Chemistry, Applied Sciences Publishers, London and New York, Vol. I, pp 83-114.
2. Pervaiz, S and Brew, K. (1985) Science 228, 335-337.
3. Ali, S. and Clark, A.J. (1988) J. Mol. Biol. 199, 415-426.
4. Simons, J.P. McClenaghan, M. and Clark, A.J. (1987) Nature 328, 530-532.
5. Harris, S. et al. in preparation.
6. Gaye, P. Hue-Delahaie, D. Mercier, J-C. Soulier, S. Vilotte, J-L. and Furet, J-P. (1986) Biochimie 68, 1097-1107.
7. Duncan, C.H. (1987) Nucl. Acids. Res. 15, 1340.
8. Watanabe, Y. Tsukada, T. Notake, M. Nakanishi, S. and Numa S. (1982) Nucl. Acids. Res. 10, 1459-1469.

Volume 16 Number 21 1988 Nucleic Acids Research

Complete nucleotide sequence of the genomic ovine β-lactoglobulin gene

S.Harris, S.Ali, S.Anderson, A.L.Archibald and A.J.Clark

AFRC Institute of Animal Physiology and Genetics Research, Edinburgh Research Station, King's Buildings, West Mains Road, Edinburgh, EH9 3JQ, UK

Submitted September 23, 1988 Accession no. X12817

ß-lactoglobulin (BLG) is the major whey protein found in the milk of a number of species (1,2). The genomic organisation of the ovine BLG gene and its relationship to a number of other secretory proteins has recently been described (3); as has the ability of sequences derived from genomic clone SS1 to direct the synthesis of large quantities of BLG protein into the milk of lactating transgenic mice (4). We present here the 7379 bps of genomic DNA sequence of clone SS1 which are sufficient to direct the tissue-specific expression of this gene in transgenic mice (5).

```
   1 gtgctcagcaacacacccagcaccagcattcccgctgctcctgaggtctgcaggcagctcgctgtagcctgagcggtgtggaggg
  86 aagtgtcctgggagatttaaaatgtgagaggcgggaggtgggaggttgggccctgtgggcctgcccatcccacgtgcctgcatta
 171 gccccagtgctgctcagccgtgcccccgccgcaggggtcaggtcactttcccgtcctggggttattatgactcttgtcattgcca
 256 ttgccatttttgctaccctaactgggcagcaggtgcttgcagagccctcgataccgaccaggtcctccctcggagctcgacctga
 341 accccatgtcacccttgccccagcctgcagagggtgggtgactgcagagatccctcacccaaggccacggtcacatggtttgga
 426 ggagctggtgcccaaggcagaggccacccctccaggacacacctgtccccagtgctggctctgacctgtccttgtctaagaggctg
 511 accccggaagtgttcctggcactggcagccagcctggacccagagtccagacacccacctgtgccccgcttctgggtctacca
 596 ggaaccgtctaggcccagaggggacttcctgcttggccttggatggaagaaggcctcctattgtcctcgtagaggaagccaccc
 681 cgggggctgaggatgagccaagtgggattccgggaaccgcgtggctgggggcccagccccgggctggctggcctgcatgcgcctcc
 766 tgtataaggccccaagcctgcctgtctcagccctccACTCCCTGCAGAGCTCAGAAGCCACGACCCCAGCTGCAGCCATGAAGTGC
 851 CTCCTGCTTGCCCTGGGCCTGGCCCTCGCCTGTGGCGTCCAGGCCATCATCGTCACCCAGACCATGAAAGGCCTGGACATCCAGA
 936 AGgttcgagggttggccgggtgggtgagttgcagggcgggcagggagctgggcctcagagagccaagagaggctgtgacgttgg
1021 gttcccatcagtcagctagggccacctgacaaatccccgctggggcagcttcaaccaggcgttcactgtcttgcattctggaggc
1106 tggaagccaagatccaggtgttggcagggctggcttctcctgcggccgctctctgggagcagacggccgtcttctccagtcct
1191 ctgcgcgccctgatttcctcttcctgtgaggccaccaggcctgctggaaacacgcctgcctgcgcagcttcacacgacctttgtc
1276 atctctttaaaggccatgtctccagagtcatgtgttgaagttctgggggttagtgggacacagttcagccccctaaaagagtctct
1361 ctgcccctcaaattttccccacctccagccatgtctccccaagatccaaatgttgctacatgtgggggggctcatctgggtccct
1446 ctttgggttcagtgtgagtctgggagagcattccccagggtgcagagttgggggagtatctcagggctgcccaggccggggtg
1531 ggacagagagcccactgtgggggctgggggcccctccccaccccagagtgcaactcaaggtccctctccagGTGGCGGGGACTTG
1616 GCACTCCTTGGCTATGGCGGCCAGCGACATCTCCCTGCTGGATGCCCAGAGTGCCCCCCTGAGAGTGTACGTGGAGGAGCTGAAG
1701 CCCACCCCCGAGGGCAACCTGGAGATCCTGCTGCAGAAATGgtgggcgtctctccccaacatggaaccccactccccagggctg
1786 tggaccccccgggggtgggtgcaggagggaccagggccccagggctgggaagagggctcagagtttactggtacccggcgct
1871 ccacccaaggctgcccacccagggctttttttttttttaaacttttattaatttgatgcttcagaacatcatcaaacaaatgaac
1956 ataaaacattcatttttgtttacttggaaggggagataaaatcctctgaagtggaaatgcatagcaaagatacatacaatgaggc
2041 aggtattctgaattccctgttagtctgaggattacaagtgtatttgagcaacagagagacattttcatcatttctagtctgaaca
2126 cctcagtatctaaaatgaacaagaagtcctggaaacgaagcagtgtggggataggcccgtgtgaaggctgctgggaggcagcaga
2211 cctgggtcttcgggctcaagcagttcccgctaccagccctgtccacctcagacgggggtcagggtgcaggagagagctggatggg
2296 tgtgggggcagagatggggacctgaaccccagggctgccttttgggggtgcctgtggtcaaggctctccctgaccttttctctct
2381 ggcttcatctgacttctcctggcccatccaccccggtccccctgtggcctgaggtgacagtgagtgcgccgaggctagttggccagc
2466 tggctcctatgcccatgccaccccccctccagcctcctgggccagcttctgcccctggcccctcagttcatcctgatgaaaatggt
2551 ccatgccaatggctcagaaagcagctgtctttcagGGAGAACGGCGAGTGTGCTCAGAAGAAGATTATTGCAGAAAAAACCAAGA
2636 TCCCTGCGGTGTTCAAGATCGATGgtgagtccgggtccctgggggacacccaccaccccgcccccgggactgtggacaggttc
2721 agggggctggcgtcgggccctgggatgctaagggactggtggtgatgaagacactgccttgacacctgcttcacttgcctcccct
2806 gccacctgcccgggccttgggcggtggccatgggcaggtcccggctggcgggctaacccaccagggtgacacccgagctctct
2891 ttgctgggggcgggcggtgctctgggccctcaggctgagctcaggaggtacctgtgccctcccaggggtaaccgagagccgttg
2976 cccactccagggggcccaggtgccccacgaccccagcccgctccacagctccttcatctcctggagacaaactctgtccgccctcg
3061 ctcattcacttgttcgtcctaaatccgagatgataaagcttcgaggggggggttggggttccatcagggctgcccttccgccgggc
3146 agcctgggccacatctgccctggcccctcaggactcactctgactggaggccctgcactgactgacgccagggtgcccagccc
3231 agggtctctggcgccatccagctgcactgggtttgggtgctggtcctgccccaagctgcccggacaccacaggcagccggggct
3316 gcccactggcctcggtcagggtgagccccagctgcccccgctcagggcttgccccgacaatgaccccatcctcaggacgcacccc
3401 ccttccctgctgggcagtgtccagccccacccgagatcgggggaagccctatttcttgacaactccagtccctgggggagggggg
3486 cctcagactgagtggtgagtgttcccaagtccaggaggtggtggagggtcctggcggatccagagttgacagtgagggcttcctg
3571 ggccccatgcgcctggcagtggcagcagggaagaggaagcaccatttcaggggtgggggatgccagaggcgctccccaccccgtc
3656 ttcgccgggtggtgaccccgggggagccccgctggtcgtggagggtgctgggggctgactagcaaccccctccccccccgttggaa
3741 ctcacttttctccctcttgaccgcgtccagCCTTGAATGAGAACAAAGTCCTTGTGCTGGACACCGACTACAAAAAGTACCTGC
3826 TCTTCTGCATGGAAAACAGTGCTGAGCCCGAGCAAAGCCTGGCCTGCCAGTGCCTGGgtgggtgccaaccctggctgcccaggga
3911 gaccagctgcgtggtccttgctgcaacaggggggtgggggggtgggagcttgatccccaggaggaggaggggtgggggtccctgag
3996 tcccgccaggagagagtggtcgcataccgggagccagtctgctgtgggcctgtgggtggctggggacgggggccagacacacagg
4081 ccgggagacgggtgggctgcagaactgtgactggtgtgaccgtcgcgatgggccggtggtcactgaatctaacagcctttgtta
```